OpenBench: a land models evaluation system

Zhongwang Wei[1*], Qingchen Xu[1], Fan Bai[1], Xionghui Xu[1], Zixin Wei[1], Wenzong Dong[1], Hongbin Liang[1], Nan Wei[1], Xingjie Lu[1], Lu Li[1], Shupeng Zhang[1], Hua Yuan[1], Laibao Liu[2,3], Yongjiu Dai[1]

[1]Southern Marine Science and Engineering Guangdong Laboratory (Zhuhai), School of Atmospheric Sciences, Sun Yat-sen University, Guangzhou, China
[2]Department of Geography, The University of Hong Kong, Hong Kong, China
[3]Institute for Climate and Carbon Neutrality, The University of Hong Kong, Hong Kong, China

Correspondence to: Zhongwang Wei (weizhw6@mail.sysu.edu.cn)

Abstract. Recent Land surface models (LSMs) have evolved significantly in complexity and resolution, requiring comprehensive evaluation systems to assess their performance. This paper introduces The **Open** Source Land Surface Model **Bench**marking System (**OpenBench**), an open-source, cross-platform benchmarking system designed to evaluate the state-of-the-art LSMs. OpenBench addresses significant limitations in the current evaluation frameworks by integrating processes that encompass human activities, facilitating arbitrary spatiotemporal resolutions, and offering comprehensive visualization capabilities. The system utilizes various metrics and normalized scoring indices, enabling a comprehensive evaluation of different aspects of model performance. Key features include automation for managing multiple reference datasets, advanced data processing capabilities, and support for station-based and gridded data evaluations. By examining case studies on river discharge, urban heat flux, and agricultural modeling, we illustrate OpenBench's ability to identify the strengths and limitations of models across different spatiotemporal scales and processes. The system's modular architecture enables seamless integration of new models, variables, and evaluation metrics, ensuring adaptability to emerging research needs. OpenBench provides the research community with a standardized, extensible framework for model assessment and improvement. Its comprehensive evaluation capabilities and efficient computational architecture make it a valuable tool for both model development and operational applications in various fields.

1 Introduction

Land surface models (LSMs) simulate the complex interactions among the land surface, planetary boundary layer, rivers and lakes, glaciers and frozen soils, plant physiology and ecology, vegetation dynamics, biogeochemistry, human activities, and other processes occurring on the land surface (Blyth et al., 2021; Dai et al., 2003; Lawrence et al., 2019; Pokhrel et al., 2016). These models play an important role in understanding and predicting various changes in the Earth system, serving as a bridge connecting the land surface, ocean, and atmosphere (Fisher and Koven, 2020; Ward et al., 2020; Liu et al., 2024). As such, they are key components of Earth system models (ESMs) and have significant impacts on our ability to comprehend and predict weather, climate, hydrological cycles, carbon cycles, and various other environmental factors. In recent decades, LSMs have undergone rapid development, evolving from basic "bucket" models (Manabe, 1969) to advanced multi-module systems

(Blyth et al., 2021) that incorporate both biogeochemical processes (e.g., greenhouse gas, carbon, nitrogen, and phosphorus cycles), geophysical processes (including land use changes, three-dimensional surface water, subsurface flow, and flooding), as well as human activities (such as agriculture, reservoir management, and urban development). This evolution has been driven by advances in hydrology, meteorology, computer science, and measurement technology, leading to the development of increasingly complex models. Concurrent with the increasing complexity of processes represented in LSMs, there has been a significant improvement in spatial resolution as well. Models have progressed from traditional forecasting scales of 25-100 km (Dai et al., 2003) to current fine scales of 0.1-10 km (Chen et al., 2024). The increasing complexity and resolution of models require a comprehensive evaluation and analysis of simulation results.

In recent years, various model benchmarking systems (See **Table 1**) have been developed. These systems assess model performance in comparison to multiple sources of reference datasets. Most of these benchmarking systems consist of benchmark datasets, evaluation software, metrics, model operating environments, and auxiliary tools. The benchmark datasets standardize observation, reanalysis, and remote sensing data to evaluate model accuracy in simulating land processes. Evaluation software includes metrics, execution environments, and tools designed for automated assessment and quantifying LSMs' performance. The operating environment comprises the software and hardware for running evaluations, while ancillary tools support benchmarking. Despite the importance of LSM evaluation and the development of various benchmarking systems, several limitations persist in current evaluation approaches. These limitations have become increasingly apparent as the complexity and resolution of LSMs have increased. One significant area for improvement is the scope of evaluation variables in most existing evaluation systems. These systems typically focus on some range of commonly used variables, such as water, heat and carbon fluxes, temperature, and vegetation coverage. This restricted scope fails to capture the full range of processes simulated by modern LSMs. For instance, TraceMe (Zhou et al., 2021) is primarily designed to evaluate model outputs related to the carbon cycle, while the MetEva software developed by the National Meteorological Center of China (https://github.com/nmcdev/meteva) focuses on meteorological fields. However, neither of these tools provides a comprehensive assessment of land surface processes, nor can they easily adapt to new evaluation indicators or datasets. In particular, there is a lack of comprehensive evaluation for hydrological cycles and human activities, making it challenging to fully assess model performance in these critical areas. Human activity, while an important factor affecting surface processes, is one of the most challenging aspects to model and evaluate. This is primarily due to the small-scale nature of human activity data (e.g., crop yields, dam operation, and anthropogenic heat) and the involvement of complex socio-economic and land-use change data. These datasets are often multi-source, complex, and highly uncertain. To date, no evaluation system has been found to integrate the assessment of human activities in LSMs broadly.

Another significant challenge in current LSMs evaluation practices is the difficulty in conducting inter-model comparisons. This comparative work is a key step in improving model performance and understanding model differences and uncertainties. However, the lack of a universal and comprehensive evaluation tool presents significant challenges, especially in the context of high-resolution complex models and evolving underlying datasets. Traditionally, software tools for evaluating LSMs have often been customized for specific models or datasets. For example, the evaluation tools for The Canadian Land Surface

Scheme (CLASS) (Verseghy, 1991) and The Community Atmosphere Biosphere Land Exchange model (CABLE) (Haverd et al., 2018), i.e., AMBER (Arora et al., 2023) and benchcab (https://github.com/CABLE-LSM/benchcab), are optimized for their respective outputs. Customized software designs lead to several issues. Researchers must invest time in learning specific usage methods and data formats for each new model, limiting their ability to try new models and slowing down comparisons. Different tools using different evaluation criteria and formats make it difficult to compare model performance. Evaluation tools like the International Land Model Benchmarking (ILAMB) platform (Collier et al., 2018) require complex data processing, such as converting model outputs to the Coupled Model Intercomparison Project (CMIP) standard. This process consumes time and computing resources, increasing the risk of errors and potentially affecting the reliability of evaluations. In the meantime, some platforms, such as ILAMB and the Land Surface Verification Toolkit (LVT) (Kumar et al., 2012), offer a wide range of assessments for process variables. However, their spatiotemporal resolution is relatively low (typically at a monthly scale and 0.5°). They have limitations in processing data conversion at different scales, making it difficult to perform simulation evaluations at multiple spatiotemporal scales.

Visual analysis capabilities are another area where current evaluation tools often fall short. Many lack visual functions or produce low-quality visualizations, making it difficult to display evaluation results effectively. For instance, while some can produce graphical diagnostics, the quality is often insufficient to meet publication standards, and it is unable to customize output. Platform compatibility is also a significant issue, as most evaluation tools are designed to run only on Linux. This limits their application on Windows or macOS operating systems, thus restricting their popularity and accessibility.

Table 1. The software that can be used or partly used for land surface model evaluation. The abbreviations for specific nouns in the table are described below:AMWG: NCAR's CAM Diagnostics Package; CVDP: NCAR's Climate Variability Diagnostics Package; ESMValTool: Earth System Model Evaluation Tool; PMP: PCMDI's Metrics Package; ILAMB: International Land Model Benchmarking System; MDTF: NOAA's Model Diagnostics Task Force Framework; MPAS-Analysis: analysis for MPAS (Model for Prediction Across Scales) components of E3SM Ocean and Sea-ice analysis for E3SM's MPAS components; E3SM Diags v2.7: The E3SM Diagnostics Package; AMBER: Automated Model Benchmarking; PALS: Protocol for the Analysis of Land Surface models; LVT: Land Surface Verification Toolkit; benchcab: Evaluation Tool for the Land Surface Model CABLE; TraceMe: Traceability analysis system for model evaluation; AMET: The Atmospheric Model Evaluation Tool; MET: The Model Evaluation Tools; MVIETool: the Multivariable Integrated Evaluation Tool.

Name	Range of application	Arbitrary spatiotemporal resolution	Cross-platform	Reference	Link
MetEva	GRAPES model	No	Yes	NA	https://github.com/nmcdev/meteva

AMWG (retired)	CAM	No	No (Linux)	NA	https://www2.cesm.ucar.edu/working_groups/Atmosphere/amwg-diagnostics-package/
CVDP	CMIP-style	No	No (Linux)	Phillips et al. (2014)	https://www2.cesm.ucar.edu/working-groups/cvcwg/cvdp
ESMValTool	CMIP-style	No	No (Linux)	Weigel et al. (2020)	https://esmvaltool.org/
PMP	CMIP-style	No	No (Linux)	Lee et al. (2023)	https://github.com/PCMDI/pcmdi_metrics
MDTF	Single point, CMIP-style, NCAR, and GFDL model	Yes	No (Linux and macOS)	NA	https://mdtf-diagnostics.readthedocs.io/en/latest/
MPAS-Analysis	MPAS model	Yes	No (Linux)	NA	https://github.com/MPAS-Dev/MPAS-Analysis
E3SM Diags	E3SM model, CMIP-style	Yes	No (Linux)	C. Zhang et al. (2022)	https://github.com/E3SM-Project/e3sm_diags
ILAMB	CMIP-style	No	No (Linux)	Collier et al. (2018)	https://www.ilamb.org/

AMBER	CLASSIC, CTEM	Yes	Yes	Arora et al. (2023)	https://cccma.gitlab.io/classic_pages/benchmarking/
modelevaluation.org	Single point	Yes	Yes	Best et al. (2015)	https://modelevaluation.org/
LVT	Various	Yes	No (Linux)	Kumar et al. (2012)	https://github.com/NASA-LIS/LISF/tree/master/lvt
benchcab	Single point, CABLE model	Yes	Yes	-	https://github.com/CABLE-LSM/benchcab
TraceME	CMIP-style	Yes	No (Linux)	Zhou et al. (2021)	http://traceme.org.cn/
AMET	CMAQ model	Yes	No (Linux)	Appel et al. (2011)	https://www.epa.gov/cmaq/atmospheric-model-evaluation-tool
MAT	WRF, UFS, and SIMA model	Yes	No (Linux)	Jensen et al. (2024)	https://metplus.readthedocs.io/projects/met/en/latest/
MVIETool	CMIP-style	No	No (Linux)	M.-Z. Zhang et al. (2021)	https://github.com/Mengzhuo-Zhang/MVIETool

95

To address these challenges and meet the high standard requirements of new-generation LSM verification and evaluation, we have developed OpenBench (The Open Source Land Surface Model Benchmarking System). The core goal of OpenBench is to provide an open-source, fast, efficient, diverse, and accurate evaluation mechanism for high-resolution land-surface model outputs. OpenBench is designed as a universal and high-performance LSM evaluation system, fully written in Python, that

100 realizes functions such as data processing, evaluation method encapsulation, and result analysis visualization. OpenBench supports cross-platform, including Windows, macOS, and Linux, enhancing its accessibility and usability across different research environments. OpenBench incorporates evaluation metrics and datasets that account for human activities on land surface processes, filling a significant gap in current evaluation systems. The system provides a unified and standardized benchmark test method framework, allowing for efficient and comprehensive validation and evaluation of typical land surface

105 models, such as CoLM, CLM, Noah-MP, GLDAS, and JULES, as well as CMIP-style model output. By ensuring the widespread sharing of evaluation results, OpenBench aims to advance scientific research and operational work in land surface modeling. The system maximizes the use of available observational and reanalysis data through its efficient data management and processing capabilities.

In the following sections of this paper, we will detail the methodology behind OpenBench, including its system architecture,

110 key components, and the benchmark datasets developed for it. We will then present case studies that demonstrate its application in evaluating and comparing different LSMs or parameterizations, highlighting its capabilities in handling high-resolution data and complex processes. Finally, we will discuss the implications of this new evaluation system for the field of land surface modeling and outline future directions for its development and application.

2 Overall Structure

115

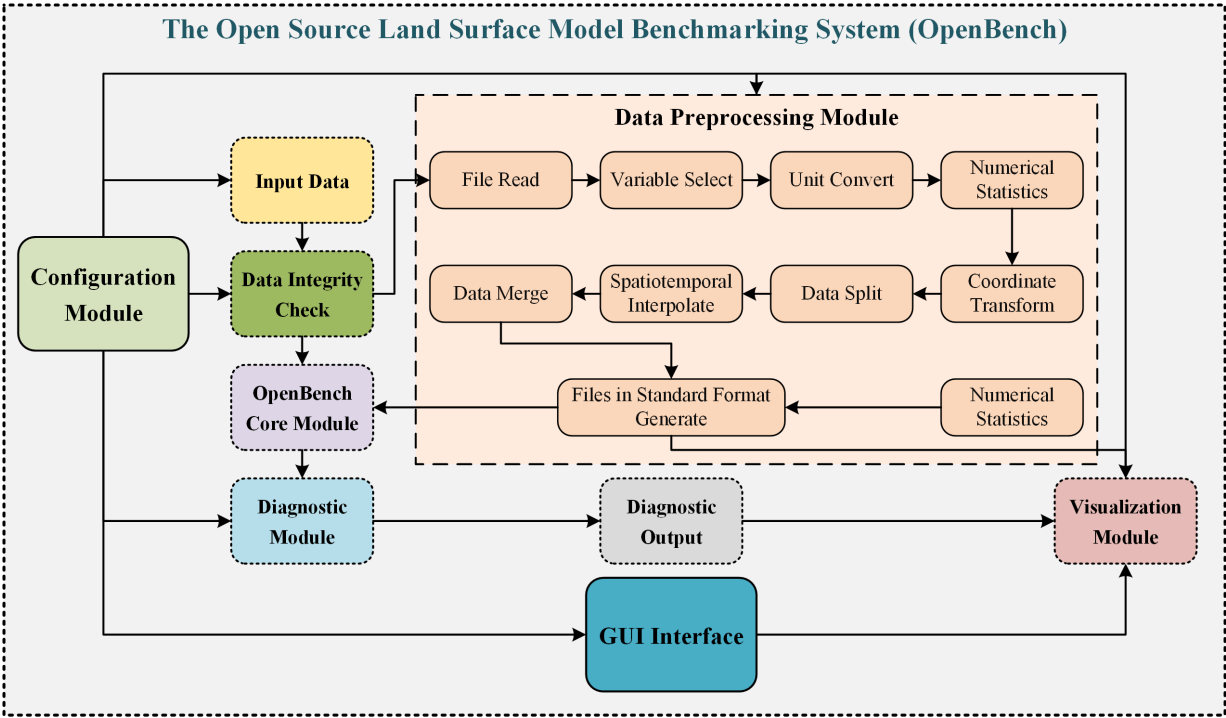



Figure 1: General flowchart of the OpenBench.

OpenBench represents a significant advancement in the field of model evaluation and intercomparison. This section outlines the system's overall structure, highlighting its key components and workflow. OpenBench is designed with modularity and flexibility in mind, enabling efficient processing of diverse datasets and model outputs. The OpenBench code is designed to simultaneously handle various data types, including plot-scale data (such as station data) and gridded data (regional or global) for both simulation and reference datasets. The flowchart of OpenBench is shown in **Fig. 1**.

The system includes six components, i.e., configuration management, data processing, evaluation, comparison processing, statistical analysis, and visualization. The configuration management module accommodates three configuration namelist formats (YAML, JSON, and Fortran namelist) to meet different user preferences and workflows, with JSON as the default format. Users can utilize the configuration namelist to define evaluation parameters, data sources, and model outputs. This adaptable configuration system facilitates straightforward customization of evaluation scenarios. The data processing module handles the preprocessing of both reference and simulation data, including temporal and spatial resampling to ensure consistent comparison between datasets with different spatiotemporal resolutions. The evaluation module implements the core evaluation logic, applying various metrics and scores to quantify model performance. It supports both gridded and station-based data and adapts its methods accordingly. The comparison module facilitates multi-model and multi-scenario comparisons, enabling comprehensive analysis across different models or configurations. Finally, advanced statistical techniques are implemented in the statistical analysis module, providing deeper insights into model behaviors and performance patterns. The system also includes capabilities for generating visualizations of evaluation results, which are crucial for interpreting and communicating findings provided in the visualization module.

The system's workflow follows a logical sequence of operations. The process begins with initialization, where command-line arguments are parsed, and configuration files are read. This stage sets up necessary directories and initializes key variables, laying the groundwork for subsequent operations. The system then moves into the data preparation phase, where both observational and model data are processed to ensure compatibility in terms of temporal and spatial resolution. This crucial step handles various data formats and structures, normalizing them for consistent analysis. At the core of the system is the evaluation process. Here, a wide array of metrics and scores is applied to quantify the agreement between model outputs and observational data. This step is highly parallelized to efficiently handle large datasets, allowing for a comprehensive assessment across multiple variables and timeframes. If multiple models or scenarios are being evaluated, the system performs comparative analyses to highlight relative strengths and weaknesses. This comparison stage provides valuable perspectives into model performance across different conditions or implementations. Following the primary evaluation, the system conducts advanced statistical analyses to gain a profound understanding from the evaluation results. This may include uncertainty quantification, trend analysis, or other sophisticated statistical methods. The final stages involve result generation and visualization. The system compiles evaluation results, generates summary statistics, and prepares data for visualization. The system can produce various charts, graphs, and maps to effectively communicate the evaluation outcomes. Throughout these stages, the system demonstrates flexibility in handling different types of data (grid-based or station-based), various temporal resolutions, and a wide range of environmental variables. It also incorporates specialized handling for different land surface models, recognizing

the unique characteristics and outputs of each. This comprehensive approach allows for a thorough, standardized evaluation of land surface models, providing valuable feedback for model development and application in Earth system science.

Nevertheless, OpenBench is developed to serve as a specialized tool for land surface model output analysis, evaluation, and comparison. The software package is freely available to the community. The code is modular and can be easily extended or modified to accommodate the specific requirements of different evaluation tasks. OpenBench relies on various popular and well-established Python packages specific to the scientific computing stack: NumPy (Harris et al., 2020), Xarray (Hoyer and Hamman, 2017), Pandas (Mckinney, 2010), SciPy (Virtanen et al., 2020), Matplotlib (Hunter, 2007), Cartopy (Met Office, 2010), Dask (Rocklin, 2015), and Joblib (Joblib, 2020). The remap functions have several options: SciPy, Cdo (Schulzweida, 2023), xesmf (Zhuang et al., 2023), and xarray-regrid (Schilperoort et al., 2024) are available for selection. We use as few packages as possible, reducing dependencies to improve performance and compatibility. The software is developed and hosted on GitHub and is distributed under the Apache-2.0 license. The latest version of OpenBench can be found in the Zenodo repository, where it has been assigned a Digital Object Identifier (https://doi.org/10.5281/zenodo.14540647).

OpenBench achieves speed improvements through its parallel processing architecture. Benchmark tests demonstrate clear advantages over sequential processing methods. In station-based evaluations, processing a single variable across 142 stations takes 3.12 minutes using single-process execution, whereas parallel processing with 48 cores reduces this to 0.509 minutes on an Intel(R) Xeon(R) CPU E5-4640 v4 @ 2.10 GHz with 48 GB RAM. OpenBench uses Dask's lazy execution and chunked arrays for efficient gridded data processing, balancing memory use and processing speed. Processing 0.25° resolution model outputs (2001-2010, monthly) against two reference datasets takes 2.302 minutes sequentially versus 1.301 minutes in parallel on the same hardware. These performance improvements are particularly beneficial for comprehensive model evaluations involving multiple variables, reference datasets, and spatial domains. The efficiency gains from parallel processing become more substantial with higher-resolution datasets and increasing numbers of evaluation sites, making OpenBench suitable for both rapid diagnostic evaluations on personal workstations and extensive comparative studies on high-performance computing systems.

3 Evaluation

3.1 The metric index

Table 2. Metrics employed in OpenBench

Metric	Full name	Range	Reference	Additional Description
Bias metrics (The smaller is better. The ideal value is 0.)				
BIAS	Bias	$(-\infty, +\infty)$	Cole (1981)	-
PBIAS	The percentage of bias	$(-\infty, +\infty)$	Sorooshian et al. (1993)	-
APBIAS	Absolute Percent Bias	$[0, +\infty)$	Sorooshian et al. (1993)	-

PC_MAX	Percent Bias of Maximum	(-∞, +∞)	X. Zhou et al. (2024)	Measuring the bias of a model when predicting the maximum value.
PC_MIN	Percent Bias of Minimum	(-∞, +∞)	X. Zhou et al. (2024)	Measuring the bias of a model when predicting the minimum value.
PC_AMPLI	Percent Bias of Amplitude	(-∞, +∞)	X. Zhou et al. (2024)	Measuring the bias of a model when predicting the data range.
APFB	Annual high flow percent bias	(-∞, +∞)	Mizukami et al. (2019)	Measuring the relative bias between simulated and observed annual peak flows
PBIAS_HF	Percent Bias of High Flows	(-∞, +∞)	Mizukami et al. (2019)	Measuring the model's bias in predicting high flows (typically above the 98th percentile).
PBIAS_LF	Percent Bias of Low Flows	(-∞, +∞)	Mizukami et al. (2019)	Measuring the model's bias in predicting low flows (typically below the 30th percentile).
PBIAS_FDC	Percent Bias in the Slope of the Mid-segment of the Flow Duration Curve	(-∞, +∞)	Yilmaz et al. (2008)	Measuring the model's bias in predicting moderate flows (typically fall within the 25th to 75th percentiles).
P-factor	Percent of simulations without the given uncertainty bounds	[0, 1]	Abbaspour et al. (2007)	Measuring the percentage of reference data falling outside the given uncertainty range.
Error metrics (The closer to 0 is better. The ideal value is 0.)				
RMSE	Root Mean Square Error	[0, +∞)	Kenney & Keeping (1962)	-
MSE	Mean Square Error	[0, +∞)	Makridakis et al. (1982)	-
ubRMSE	Unbiased Root Mean	[0, +∞)	Entekhabi et al. (2010)	Remove systematic bias

	Square Error			from RMSE and only consider random errors.
CRMSE (NRMSE)	Centralized Root Mean Square Error (Normalized Root Mean Square Error)	[0, +∞)	Stephen & Kazemi (2014)	Measuring the random component of model error, independent of their mean values.
MAE	Mean Absolute Error	[0, +∞)	Yapo et al. (1996)	Less sensitive to outliers.
RSS	Residual sum of squares	[0, +∞)	Archdeacon (1994)	-
RSR	RMSE minus Observations Standard Deviation Ratio	[0, +∞)	Legates & McCabe Jr (1999)	-
IPE	The Ideal Point Error	[0, 1]	Dawson et al. (2012)	-
Correlation metrics (The larger is better. The ideal value is 1.)				
R	Correlation Coefficient	[-1, 1]	Pearson (1920)	-
R^2	Coefficient of Determination	[0, 1]	Box (1966) and Barrett (1974)	-
ubR	Unbiased Correlation Coefficient	[-1, 1]	Olkin & Pratt (1958)	Not affected by systematic bias.
ubR^2	Unbiased Coefficient of Determination	[-1, 1]	Olkin & Pratt (1958)	Not affected by systematic bias.
rSpearman	Spearman's Rank Correlation Coefficient	[-1, 1]	Spearman (1961)	Measuring the monotonic relationship between two variables.
br^2	R-squared multiplied by regression slope	[0, 1]	Krause et al. (2005) and Krstic et al. (2016)	Combines the model's bias and goodness of fit.
Efficiency metrics (The larger is better. The ideal value is 1.)				
NSE	Nash-Sutcliffe Efficiency	(-∞, 1]	Nash & Sutcliffe (1970)	-
LNSE	Log Nash-Sutcliffe Efficiency	[0, 1]	Lamontagne et al. (2020)	More sensitive to lower values.
mNSE	Modified Nash-Sutcliffe	[0, 1]	Legates & McCabe Jr (1999)	Using absolute differences

	Efficiency			instead of squared differences.
rNSE	Relative Nash-Sutcliffe Efficiency	(-∞, 1]	Legates & McCabe Jr (1999)	Suitable for evaluating relative errors.
wsNSE	Weighted Seasonal Nash-Sutcliffe Efficiency	(-∞, 1]	Zambrano-Bigiarini & Bellin (2012)	Allows for evaluating model performance across different seasons while considering the relative importance of seasons.
KGE	Kling-Gupta Efficiency	(-∞, 1]	Gupta et al. (2009)	-
KGESS	Standardized Kling-Gupta Efficiency	(-∞, 1]	Knoben et al. (2019)	A normalized version of KGE, facilitating comparison between different models.
ubKGE	Unbiased Kling-Gupta Efficiency	(-∞, 1]	Gupta et al. (2009)	Removing bias calculation.
KGEkm	Kling-Gupta Efficiency with Known Moments	(-∞, 1]	Pizarro & Jorquera (2024)	Considering the Coefficient of Variation.
KGElf	Kling-Gupta Efficiency for Low Flows	(-∞, 1]	Garcia et al. (2017)	Evaluating the model's ability to predict low flows.
Agreement metric (The larger is better. The ideal value is 1.)				
IA	Index of Agreement	[0, 1]	Krause et al. (2005)	-
RIA	Relative Index of Agreement	[0, 1]	Krause et al. (2005)	-
ReIA	The Refined Index of Agreement	[0, 1]	Willmott et al. (2012)	-
valindex	Valid Index	[0, 1]	Criss & Winston (2008)	Measuring the proportion of model predictions falling within an acceptable range.
L	Likelihood Estimation	[0, 1]	Myung (2003)	Evaluating the probability of model predictions.
Others				

rSD	Ratio of Standard Deviations	(-∞, +∞)	Everitt & Skrondal (2010)	Greater than 1 indicates that the simulation has larger variability, vice versa.
RV	Relative Variability	(-∞, +∞)	Everitt & Skrondal (2010)	-
CV	Coefficient of Variation	(-∞, +∞)	Lovie (2005)	-
KC	Cohen's Kappa coefficient	[-1, 1]	Cohen (1960)	Evaluates agreement for categorical data while accounting for chance agreement.

OpenBench uses a variety of metrics to evaluate LSM performance thoroughly (**Table 2**). This approach offers different viewpoints on model behavior, detailed comprehension of model strengths and weaknesses, versatile comparison abilities for
180 both individual and inter-model assessments, and efficient implementation using Xarray and Dask software for handling large datasets. The system incorporates various categories of metrics to capture different aspects of model performance. For example, Bias metrics, such as Percent Bias (PBIAS) and Absolute Percent Bias (APBIAS), measure systematic over- or under-estimation and bias magnitude, respectively. Error metrics, including Root Mean Square Error (RMSE), Unbiased Root Mean Square Error (ubRMSE), Centralized Root Mean Square Error (CRMSE), and Mean Absolute Error (MAE), provide different
185 perspectives on the magnitude and nature of model errors. Efficiency metrics like Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta Efficiency (KGE) evaluate model performance relative to baselines and combine multiple aspects of the model-data agreement. Correlation metrics, including Pearson correlation coefficient (R) and coefficient of determination (R^2), quantify the strength and direction of linear relationships between model outputs and observations. The Index of Agreement (IA) provides a more comprehensive assessment of magnitude and phase agreement. Variability metrics such as Ratio of Standard
190 Deviations (rSD) and specialized bias metrics for maximum (PC_MAX), minimum (PC_MIN), and amplitude (PC_AMPLI) values help identify whether models accurately capture the range of system variability and extreme conditions. For categorical data, the Cohen's Kappa coefficient (KC) evaluates agreement while accounting for chance. Variability metrics such as Relative Variability (RV) and Coefficient of Variation (CV) help identify whether models accurately capture the range of system variability. Bias-corrected versions of several metrics focus on assessing agreement in variability patterns after
195 removing mean biases. In summary, this comprehensive approach provides a robust foundation for quantitative LSM's assessment, enabling a multi-faceted evaluation that captures various aspects of the model-observation agreement. By implementing this range of metrics, OpenBench offers a thorough and nuanced evaluation of LSMs, supporting scientific understanding and practical model improvement.

3.2 The scoring index

200   OpenBench implements a suite of normalized score indices developed in ILAMB (Collier et al., 2018; Arora et al., 2023), ranging from 0 to 1, with 1 indicating perfect agreement between the model and observations. ILAMB encompasses several key indices, each designed to evaluate specific aspects of model performance. The Normalized Bias Score (nBiasScore) quantifies systematic errors in the model's predictions, normalized by observational variability. For a given variable $v(t, x)$, where t represents time and x represents spatial coordinates, we first calculate the bias from the temporal means of both the

reference $\overline{v_{ref}}(x)$ and model $\overline{v_{sim}}(x)$ data. To score the bias, we normalize it by the Centralized Root Mean Square (CRMS) of the reference data:

$$CRMS(x) = \sqrt{\frac{\int_{t_0}^{t_f} \left(v_{ref}(t,x) - \overline{v_{ref}}(x)\right)^2 dt}{t_f - t_0}} \qquad (1)$$

Where t_0 and t_f are the first and final timestep, respectively. We then compute the bias, $bias(x) = v_{ref}(t,x) - v_{sim}(t,x)$. The relative error in bias is then given as $\varepsilon_{bias}(x) = \frac{|bias(x)|}{CRMS(x)}$. The bias score as a function of space is then computed as:

$$nBiasScore(x) = e^{-\varepsilon_{bias}(x)} \qquad (2)$$

This score effectively penalizes large biases relative to the natural variability of the system. To evaluate the model's ability to capture observational variability, we employ the Normalized RMSE Score (nRMSEScore). Similar to nBiasScore, we first calculate the Centralized RMSE (CRMSE):

$$CRMSE(x) = \sqrt{\frac{\int_{t_0}^{t_f} \left(\left(v_{sim}(t,x) - \overline{v_{sim}}(x)\right) - \left(v_{ref}(t,x) - \overline{v_{ref}}(x)\right)\right)^2 dt}{t_f - t_0}} \qquad (3)$$

The relative error in bias is then given as $\varepsilon_{rmse}(x) = \frac{CRMSE(x)}{CRMS(x)}$. The nRMSEScore as a function of space is then computed as:

$$nRMSEScore(x) = e^{-\varepsilon_{rmse}(x)} \qquad (4)$$

This metric is particularly sensitive to differences in variability patterns between model outputs and observations. For variables with strong seasonal patterns, the Normalized Phase Score (nPhaseScore) assesses the model's ability to capture the timing of seasonal cycles, providing insight into the model's representation of temporal dynamics. The nPhaseScore is calculated as:

$$nPhaseScore(x) = \frac{1}{2}\left[1 + cos\left(\frac{2\pi\theta(x)}{nstep}\right)\right] \qquad (5)$$

where $\theta(x, \lambda, \phi)$ is the time difference between modeled and observed maxima:

$$\theta(x) = maxima\left(c_{sim}(x,t)\right) - maxima\left(c_{ref}(x,t)\right) \qquad (6)$$

Here, c_{sim} and c_{ref} are the climatological mean cycles (i.e., the average seasonal patterns) of the model and reference data, computed by averaging each month or day across all years in the time series. The "*maxima*" (   ) identifies the timing (month for monthly data, day for daily data) when the peak value occurs in these average seasonal cycles at each spatial location x. The parameter *nstep* represents the number of time steps in a complete annual cycle (e.g., 12 for monthly data or 365 for daily data) and normalizes the phase difference to the annual cycle.

Interannual variability, a critical aspect of climate modeling, is evaluated using the Normalized Interannual Variability Score (nIavScore). nIavScore is given by first removing the annual cycle from both the reference and model:

$$iav_{ii}(x) = \sqrt{\frac{\int_{t_0}^{t_f}\left(v_{ii}(t,x) - c_{ii}(t,x)\right)^2 dt}{t_f - t_0}} \tag{7}$$

where ii represents sim or ref. Then, the relative error is calculated as:

$$\varepsilon_{iav}(x) = \frac{iav_{sim}(x) - iav_{ref}(x)}{iav_{ref}(x)} \tag{8}$$

Similar to Equation (2) and (4), the nIavScore is given by

$$\text{nIavScore}(x) = e^{-\varepsilon_{iav}(x)} \tag{9}$$

This score is crucial for assessing the model's performance in representing year-to-year variations driven by climate factors. The Spatial Score (nSpatialScore) evaluates how well the model captures the spatial distribution of a variable compared to observations by assessing both the spatial correlation and the relative variability across the domain. The nSpatialScore is calculated as:

$$\text{nSpatialScore} = 2(1 + R)/\left(\sigma + \frac{1}{\sigma}\right)^2 \tag{10}$$

where R is the spatial correlation coefficient between the model and reference period mean values, and σ is the ratio of spatial standard deviations:

$$\sigma = stdev\left(\overline{v_{sim}}(x)\right)/stdev\left(\overline{v_{ref}}(x)\right) \tag{11}$$

To provide an overall assessment of model performance, we calculate an Overall Score (OvScore) that combines these individual metrics. This composite score gives double weight to the nRMSEScore due to its importance in capturing both bias

245 and variability aspects, which is consistent with ILAMB. The Relative Score (ReScore) is designed to compare performance across simulations by normalizing a model's overall score relative to the multi-simulation mean and standard deviation. Positive values indicate above-average performance, while negative values indicate below-average performance. Detailed information can be obtained from Collier et al. (2018) and Arora et al. (2023).

ILAMB and OpenBench exhibit two key differences in their scoring methodologies. The first distinction lies in their approach
250 to calculating global mean scores. ILAMB applies mass weighting when evaluating variables that represent carbon or water mass/flux, such as Gross Primary Production (GPP) or precipitation. This method can lead to global mean scores being disproportionately influenced by middle and low latitudes, as exemplified by the significant impact of GPP or precipitation in the Amazon. In contrast, OpenBench offers greater flexibility in its weighting methods. OpenBench supports multiple weighting options that users can select based on their requirements. Users can choose between a simple spatial integral for
255 unweighted averaging, area weighting to account for varying grid cell sizes across latitudes, or mass weighting for mass/flux variables. This flexibility allows researchers to choose the best weighting method for their particular analysis. For example, when evaluating GPP, researchers might opt for mass weighting to align with ILAMB's methodology, or they could choose area weighting to ensure more balanced representation across latitudes. The choice of weighting method can significantly impact the final results, particularly when analyzing variables with strong spatial heterogeneity. The second major difference
260 pertains to how these systems handle multiple reference datasets. ILAMB combines evaluation results from different reference datasets, assigning weights to each and combining them multiplicatively to produce a single final score that incorporates all datasets. OpenBench, on the other hand, provides users with multiple reference datasets and allows them to select one or more that they consider most accurate. It then reports scores separately for each chosen dataset without applying weights. This approach gives users more flexibility and transparency in interpreting results, allowing them to make informed decisions based
265 on their knowledge of dataset quality and relevance to their specific research questions. These methodological differences reflect the distinct philosophies and goals of each system. ILAMB's approach emphasizes a comprehensive, weighted assessment that accounts for the relative importance of different regions and datasets. OpenBench prioritizes user choice and equal spatial representation in its scoring methodology, allowing for a more customizable and potentially more equitable evaluation process. Both approaches have their merits, and the choice between them may depend on the specific needs and
270 preferences of the research community using these benchmarking tools. It is worth noting that, although we refer to OpenBench as a "benchmarking system" in accordance with community convention, the tool primarily functions as an evaluation and comparison framework rather than adhering to strict benchmarking with predetermined performance standards. This design

choice affords users the flexibility to establish their own performance criteria while benefiting from standardized evaluation methodologies.

275 In summary, by combining multiple normalized scores that assess different aspects of model performance, we enable a nuanced understanding of model strengths and weaknesses. This approach not only supports the evaluation of individual models but also facilitates inter-model comparisons and the tracking of model improvements over time.

3.3 Datasets

OpenBench integrates a diverse array of benchmarking data spanning multiple variables, levels, and spatiotemporal resolutions.
280 This approach ensures a thorough evaluation of modern high-resolution LSMs, which require increasingly detailed and accurate input data to capture complex land-atmosphere interactions.

The strength of OpenBench lies in its extensive collection of baseline datasets, categorized into five main groups: radiation and energy cycle, ecosystem and carbon cycle, hydrology cycle, parameters and atmospheric forcing, and human activity. These datasets are derived from five primary sources: field observations, satellite remote sensing, reanalysis data, machine
285 learning, and model outputs. Each source offers unique advantages, contributing to a more comprehensive understanding of land surface processes. Field observations provide high-accuracy, ground-truth data essential for model validation and calibration. While often limited in spatial coverage, these datasets offer unparalleled accuracy and temporal resolution. Satellite remote sensing delivers extensive spatial coverage and consistent temporal sampling, which are crucial for monitoring large-scale land surface processes. Reanalysis data combines model simulations with observations to create consistent, gridded data
290 products, particularly useful for long-term studies or filling observational gaps. Model outputs and machine learning, while not direct observations, provide estimates of variables that are challenging to measure directly.

The spatial-temporal scope of OpenBench's datasets is another critical feature. Many datasets span several decades, allowing for the evaluation of long-term trends and interannual variability. This extended temporal coverage enables assessing LSMs' performance over long historical periods. The resolution ranges from coarse (e.g., 0.5° for ILAMB datasets (Collier et al.,
295 2018)) to very fine (e.g., 500m for MODIS-based products (Varquez et al., 2021)), making it possible to evaluate LSMs across different spatial scales, from global assessments to regional or plot-scale studies.

A unique aspect of OpenBench is its inclusion of datasets focused on human impacts on land surface processes. This approach recognizes the growing importance of anthropogenic factors in shaping the Earth system. Datasets like AH4GUC (Varquez et al., 2021), which provides global anthropogenic heat flux data, and GDHY (Iizumi and Sakai, 2020), offering detailed
300 information on global crop yields, enable the evaluation of urban heat island effects and agricultural impacts in LSMs.

OpenBench's inclusion of multiple datasets for each variable allows for a more robust evaluation of LSMs. This multi-dataset approach enables users to assess model performance against a range of reference data, providing a more comprehensive evaluation. For instance, in evaluating evapotranspiration, OpenBench includes datasets like GLEAM4.2a (Miralles et al., 2011), FLUXCOM (Jung et al., 2019), X-BASE (Nelson et al., 2024), Xu2024 (Xu et al., 2025), and ERA5-Land (Muñoz-

305 Sabater et al., 2021), each with its own methodology and characteristics. Users can assess model performance across multiple variables simultaneously, identifying potential compensating errors or cross-variable inconsistencies that might be missed when evaluating single variables in isolation. This multi-dimensional approach provides a more complete picture of model performance and helps guide future model development efforts. Meanwhile, OpenBench's dataset collection is designed to be expandable and updatable, ensuring its relevance in the rapidly evolving field of Earth system science. As new datasets become

310 available or existing datasets are updated, they can be seamlessly integrated into the OpenBench framework.

It is noted that while OpenBench integrates with a comprehensive collection of datasets, we cannot directly provide specific data due to copyright restrictions and licensing agreements. However, to ensure transparency and reproducibility, we have included relevant links to the original data sources in **Tables S1-S5**. These links will guide users to the appropriate platforms to access the datasets following the respective terms and conditions set by the data providers. To demonstrate the functionality

315 and structure of OpenBench, we have included a set of self-generated sample data. This sample data mimics the characteristics and format of the actual datasets, allowing users to familiarize themselves with the OpenBench framework and its capabilities without infringing on any copyright issues. We encourage users to utilize these sample datasets for initial testing and exploration of the OpenBench system and then proceed to acquire the complete datasets from the original sources for comprehensive model evaluations.

320

3.4 Supporting models

OpenBench has been designed to accommodate a diverse array of land surface models, facilitating comprehensive intercomparison and evaluation studies. This multi-model support is a key feature of OpenBench, enabling researchers to assess and compare the performance of various models across different land surface processes and variables. Currently,

325 OpenBench supports various state-of-the-art land surface models, including multiple versions of the Common Land Model (CoLM2014 and CoLM2024) (Bai et al., 2024; Dai et al., 2003; Fan et al., 2024), the Community Land Model Version 5 (CLM5) (Lawrence et al., 2019), Noah-MP 5.0 (He et al., 2023), the Minimal Advanced Treatments of Surface Interaction and Runoff model (Version 2021) (Nitta et al., 2014), Atmosphere-Vegetation Interaction Model (AVIM) (Li et al., 2002), the Global Land Data Assimilation System (GLDAS2) (Rodell et al., 2004), Today's Earth (TE)

330 (https://www.eorc.jaxa.jp/water/index.html) and the Variable Infiltration Capacity (VIC) Model (Hamman et al., 2018) and so on. OpenBench has expanded its capabilities to support Lambert Conformal projections outputs from regional climate models such as the Climate extension of the Weather Research and Forecasting model (CWRF) (Liang et al., 2012) and the Weather Research and Forecasting (WRF) model (Lo et al., 2008).

Furthermore, OpenBench supports CMIP-style series outputs, such as CMIP-style simulation, such as LS3MIP (Van Den Hurk

335 et al., 2016) and ISIMIP (Wartenburger et al., 2018), allowing for seamless integration of global climate model data into the evaluation framework. Each supported model is integrated into the system through a dedicated namelist file that maps the

model's output variables to standardized variables used within OpenBench. This approach ensures consistent comparison and evaluation across different models, regardless of their native output format or projection.

3.5 Case studies

340 To illustrate the analytical capabilities of our evaluation system, we present comprehensive case studies focusing on two critical aspects of hydrological modeling: river discharge evaluation and inundation fraction assessment. These analyses were conducted using simulations from the CaMa-Flood Version 4.22 model (Yamazaki et al., 2013), driven by 0.25° remapped daily runoff data from the Global Reach-level Flood Reanalysis (GRFR) (Yang et al., 2021). The evaluation was performed globally with 0.25° spatial resolution, utilizing observational data from the Global Runoff Data Centre (GRDC) for discharge

345 validation and the Global Inundation Extent from Multi-Satellites (GIEMS) (Prigent et al., 2020) for inundation fraction assessment. Our analysis demonstrates the system's versatility in conducting site-specific and global-scale evaluations. **Figure 2a** presents a detailed comparison of simulated and observed discharge hydrographs for a representative station. This station-level analysis reveals the model's strong performance in capturing both the magnitude and temporal variability of discharge patterns. The close alignment between simulated and observed values indicates robust model performance at the local scale.

350 **Figure 2b** illustrates the spatial distribution of KGESS values for simulated discharge across the globe. The analysis reveals distinct regional patterns in model performance. The model demonstrates particularly strong capabilities in simulating discharge across wet regions, including the Amazon basin, Japan, and the Eastern United States. However, performance metrics indicate lower accuracy in the Western United States, Central Australia, and Southern Africa. These regional variations can be attributed to several factors, including the influence of human activities, uncertainties in precipitation datasets, and

355 limitations in model parameterization schemes (Wei et al., 2020). The impact of human activities on model performance is particularly evident in regions like the Western United States, where discharge patterns are significantly modified by dam operations (Hanazaki et al., 2022). This finding underscores the importance of incorporating human water management practices in regions with intensive anthropogenic influence to achieve reliable simulation results. **Figure 2c** presents global patterns of correlation coefficients for simulated inundation fraction. The results indicate strong model performance in low-

360 latitude regions, particularly in the Amazon basin and South Asia. However, significant discrepancies emerge in high-latitude areas (above 60°N). This spatial pattern of model performance highlights the need for improved representation of snow-related processes and precipitation phase partitioning in these regions (Jennings et al., 2018).

OpenBench implements automated grouping of metrics and scores according to both IGBP and PFT classifications to provide a comprehensive evaluation of model performance across diverse ecological zones. **Figure 3** presents a detailed heatmap

365 visualization of performance indices categorized by IGBP land cover types, based on CoLM2024 simulations evaluated against X-BASE reference data (Nelson et al., 2024) for 2002-2003. The analysis incorporates six fundamental performance scores developed within the ILAMB framework, as detailed in **Sect. 3.2**. The visualization reveals several significant patterns in model performance across different ecosystems. The overall nPhaseScore of 0.84 demonstrates the model's robust capability in capturing seasonal variations across all biomes. Particularly noteworthy is the model's exceptional performance in forest

370 ecosystems, where Evergreen Needleleaf Forests (ENF), Deciduous Needleleaf Forests (DNF), and Mixed Forests (MF) exhibit consistently high nPhaseScore. These results indicate the model's sophisticated ability to simulate the complex dynamics of multi-layered forest ecosystems. However, the analysis also identifies specific challenges in certain environmental contexts. The model's performance notably decreases in extreme environments, with lower scores across multiple metrics for Snow and Ice (SNO) and Barren or Sparsely Vegetated (BSV) regions. It is important to highlight that the GPP values found

375 in SNO and BSV classes could stem from spatial or temporal misalignments between the IGBP land cover classification and GPP datasets. Specifically, pixels identified as non-vegetated during the land cover survey might have had vegetation during the GPP measurement periods, or mixed pixels may consist of minor vegetated fractions within primarily barren regions. Additionally, Evergreen Broadleaf Forests (EBF) show particularly low nBiasScore, reflecting substantial magnitude discrepancies between simulated and observed values. This finding underscores the persistent challenges in accurately

380 modeling these data-sparse, highly dynamic ecosystems. These insights have important implications for model application across different research contexts. Researchers focusing on temperate and boreal forest ecosystems can proceed with high confidence in the model's capabilities. However, studies targeting arid regions, snow-covered areas, or tropical rainforests should incorporate additional validation steps and exercise greater caution in interpreting results. This systematic evaluation across biomes thus provides essential guidance for appropriate model application in diverse ecological settings.

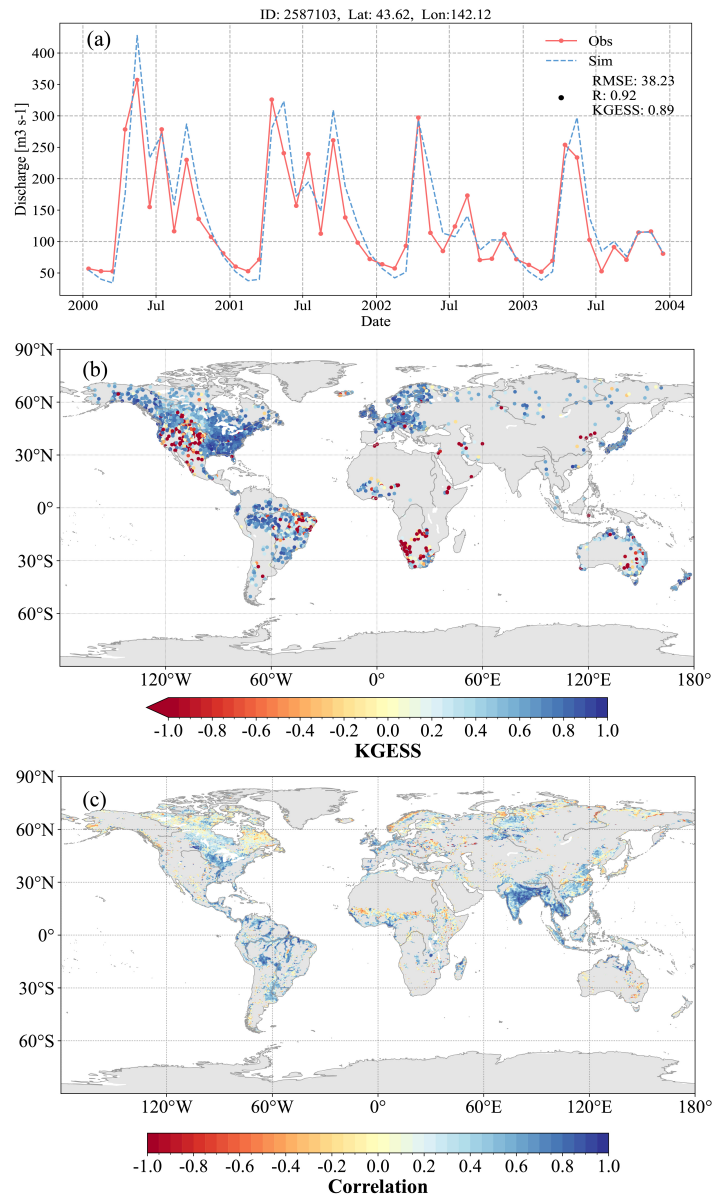



Figure 2: Example of river discharge evaluation: (a) simulated and observed discharge hydrographs for an example station; (b) global maps of KGESS values for the simulated discharge dataset; and (c) global maps of R values for the simulated inundation dataset.

385

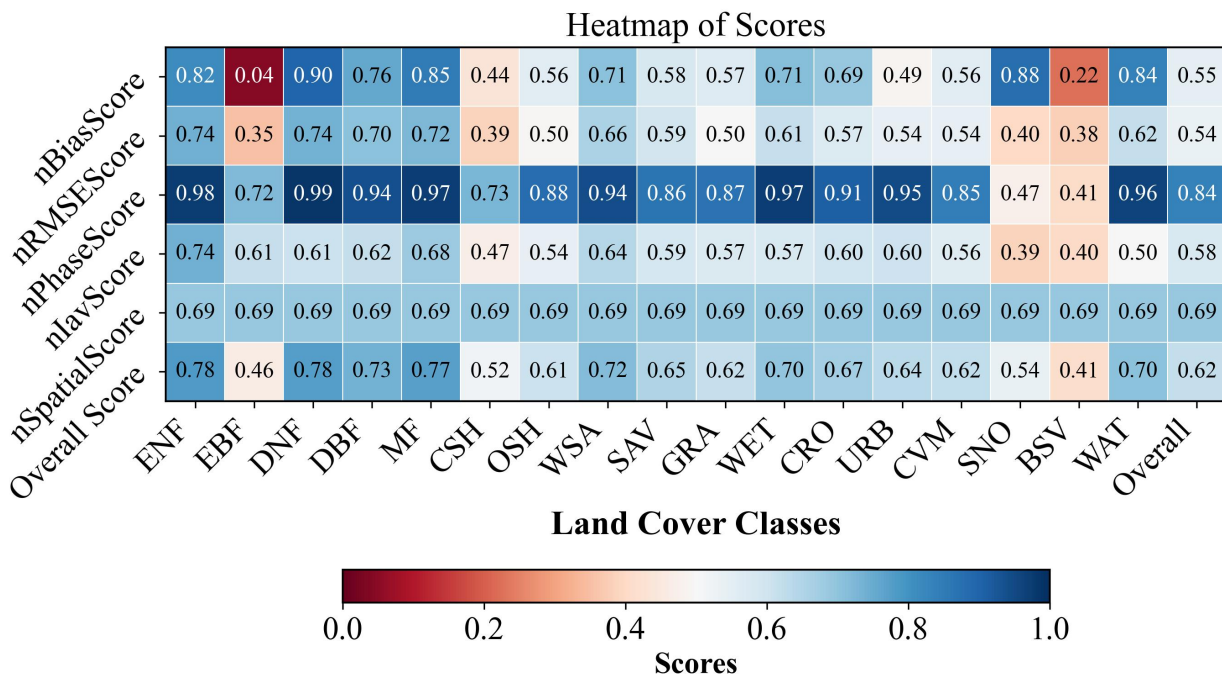



Figure 3: An example of a scores heatmap for GPP classified by IGBP land cover.

Figure 4 demonstrates OpenBench's capability to evaluate anthropogenic influences on urban thermal environments through a detailed comparison of CoLM2024 simulations with AH4GUC observational data for Southeast Asia. The analysis reveals generally strong agreement between simulated and observed anthropogenic heat flux patterns across most regions. However, notable discrepancies emerge in specific areas, particularly the corridor extending from central China to northern Vietnam and regions near Laos, where negative correlations indicate potential systematic biases in model representation. While the exact mechanisms driving these regional differences are still being investigated, these results demonstrate OpenBench's ability to identify spatial patterns of model-observation disagreement that require further exploration. The system's evaluation capabilities extend beyond thermal processes to encompass multiple aspects of human-environment interactions. Through a comprehensive assessment of variables, including latent heat, albedo, and surface temperature changes, OpenBench provides valuable insights into the complex relationships between anthropogenic activities and land surface processes, guiding improvements in their model representation.

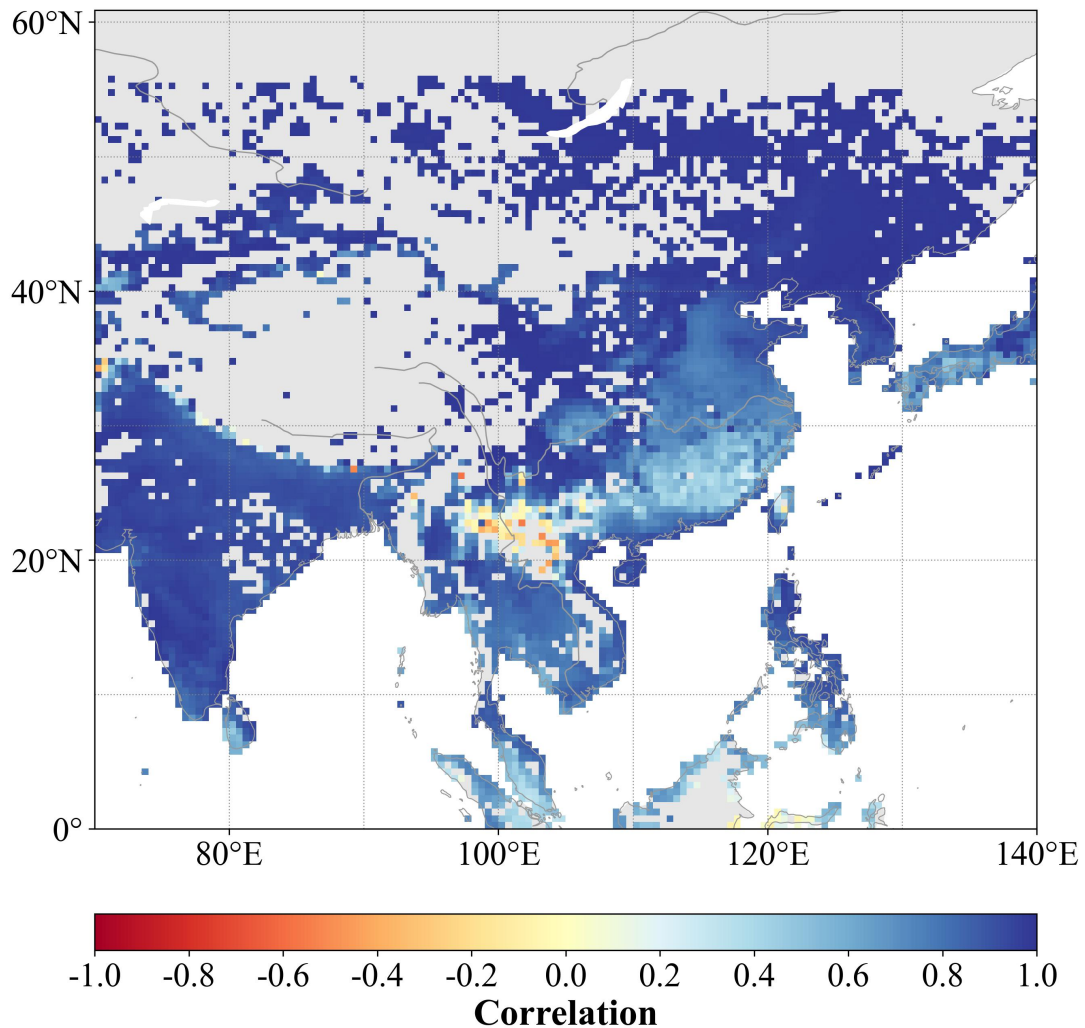



Figure 4: Pearson's correlation coefficient between CoLM2024 simulation and AH4GUC generation for urban anthropogenic heat flux over Southeast Asia.

Figure 5 presents a detailed analysis of agricultural modeling capabilities, comparing CoLM2024 simulated corn yields with GDHY-generated observational data across the United States. The analysis reveals distinct regional patterns in model performance: approximately 20% yield underestimation in the western United States, significant overestimation in central regions, and notable underestimation in eastern areas. These spatial patterns of bias may stem from multiple sources, including uncertainties in the GDHY observational dataset and the CoLM2024 model structure. Particularly noteworthy are the substantial differences in planted area and crop distribution between the two datasets, indicating fundamental challenges in representing agricultural systems within current modeling frameworks. These findings underscore the significant opportunities for advancement in both modeling and observational approaches to crop yield estimation. Future research efforts should focus

on reducing uncertainties in simulation and observational datasets while improving the representation of agricultural processes in land surface models.

In summary, these case studies demonstrate the comprehensive analytical capabilities of our evaluation system. Through its ability to conduct detailed analyses across multiple spatial scales and variables, OpenBench provides researchers with powerful tools for assessing model performance and identifying specific areas for improvement. This multi-scale, multi-variable approach supports theoretical understanding and practical application of land surface models, ultimately contributing to enhanced representation of Earth system processes.

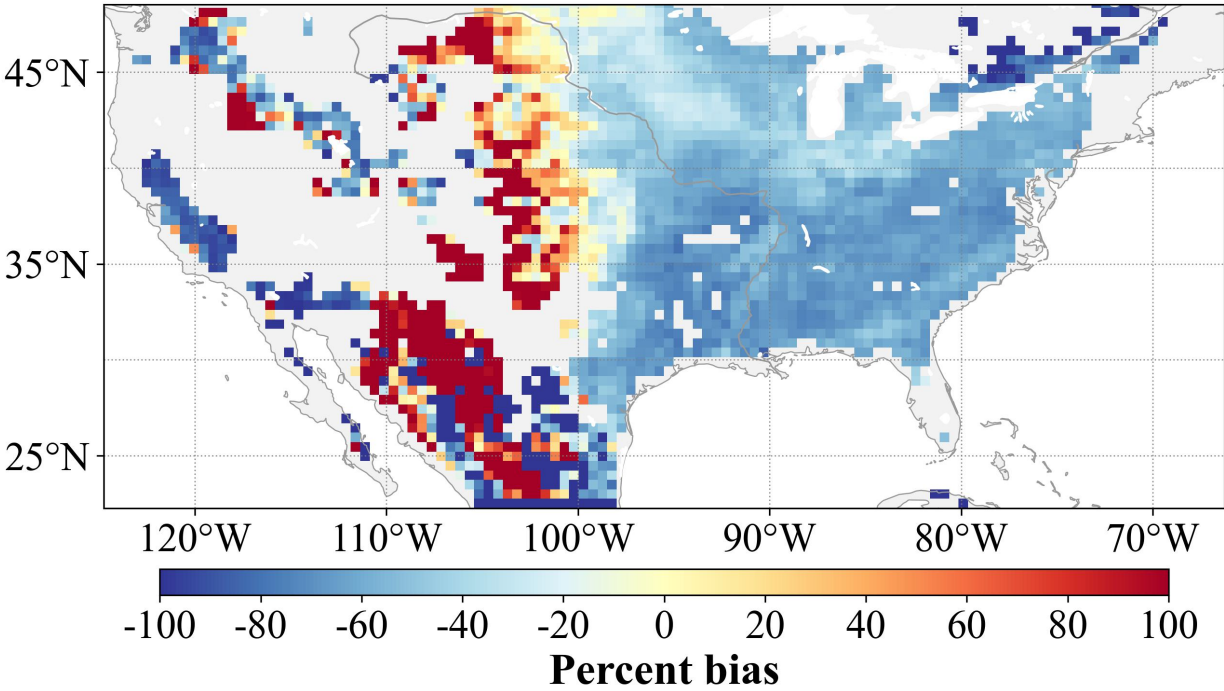



Figure 5: Percentage bias between CoLM2024 simulated and GDHY generated crop yield of corn for the United States.

4 Comparisons

4.1 Overview

OpenBench offers a comprehensive suite of comparison capabilities designed to facilitate a thorough evaluation of model performance across diverse scenarios, land cover types, and temporal scales. The system incorporates several key functionalities that enable sophisticated analysis while maintaining user accessibility and scientific rigor.

The framework's evaluation architecture encompasses multiple complementary approaches to model assessment. At its foundation, ecosystem-based comparisons allow researchers to evaluate performance across different IGBP and PFT land

cover classifications, providing crucial insights into model behavior within specific ecological contexts. This capability is enhanced by multi-metric visualization tools, including heat maps, Taylor diagrams, and target diagrams, which offer intuitive yet comprehensive overviews of model capabilities by simultaneously displaying multiple statistical metrics for model-observation comparisons. To support detailed analysis of model behavior, OpenBench implements advanced distribution and pattern analysis tools. These include kernel density estimation plots and parallel coordinate plots, which facilitate the comparison of metric distributions across models and enable the identification of patterns in multivariate performance data. The system's temporal performance evaluation capabilities, implemented through seasonal portrait plots, provide detailed insights into variations in model accuracy across different seasonal cycles. Statistical analysis within OpenBench is supported by robust summary tools, including box and whisker plots that offer concise yet comprehensive overviews of model performance across different metrics and scenarios. This statistical framework ensures that comparisons remain objective and scientifically sound while presenting results in an accessible format for interpretation.

The implementation of multiple model comparisons follows a systematic and efficient approach. The process begins with the standardization of model outputs through a sophisticated data processing pipeline, capable of handling various input formats and temporal/spatial resolutions. The comparison processing module orchestrates this analysis through support for multiple comparison methods, with parallel processing capabilities implemented via the Joblib library to ensure computational efficiency. Evaluation items and reference sources systematically organize results from the comparison process within a structured output directory. The system automatically generates comparison artifacts, including metrics and score files, which form the basis for comprehensive visualization and analysis. This structured approach ensures that adding new models to the comparison framework requires minimal effort, typically involving only the update of simulation namelist with new model information and data sources. This integrated approach to model comparison and evaluation provides researchers with powerful tools for understanding model behavior while maintaining the flexibility needed to address diverse research questions in land surface science. The system's design philosophy emphasizes scientific rigor and practical utility, ensuring that comparative analyses can be conducted efficiently while maintaining the highest standards of scientific validity.

4.2 Case studies

To demonstrate the comprehensive capability of our evaluation system, we present several case studies to demonstrate the ability of the evaluation system to compare between models, compare between different parameterized schemes, and compare between CMIP-style datasets. It's important to note that our primary goal is to showcase the evaluation system's functionality rather than to make definitive judgments about any particular model's performance. These case studies are practical examples of the system's versatility and analytical power.

 4.2.1 Multiple models comparison

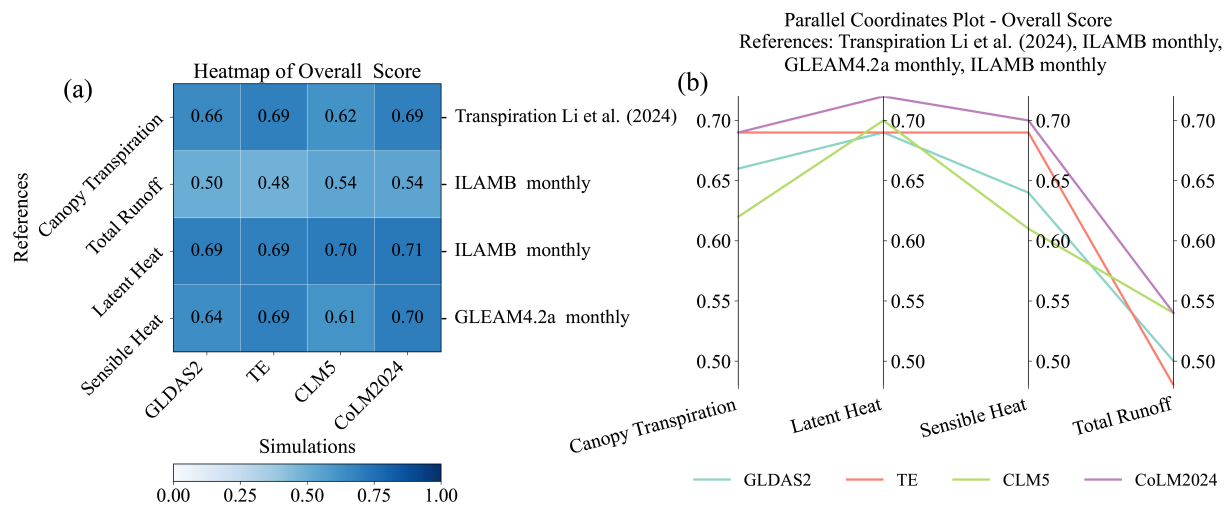



Figure 6: Overall score comparisons of sensible heat, latent heat, total runoff, and canopy transpiration using (a) heat map and (b) parallel coordinates approaches for GLDAS2, TE, CLM5, and CoLM2024.

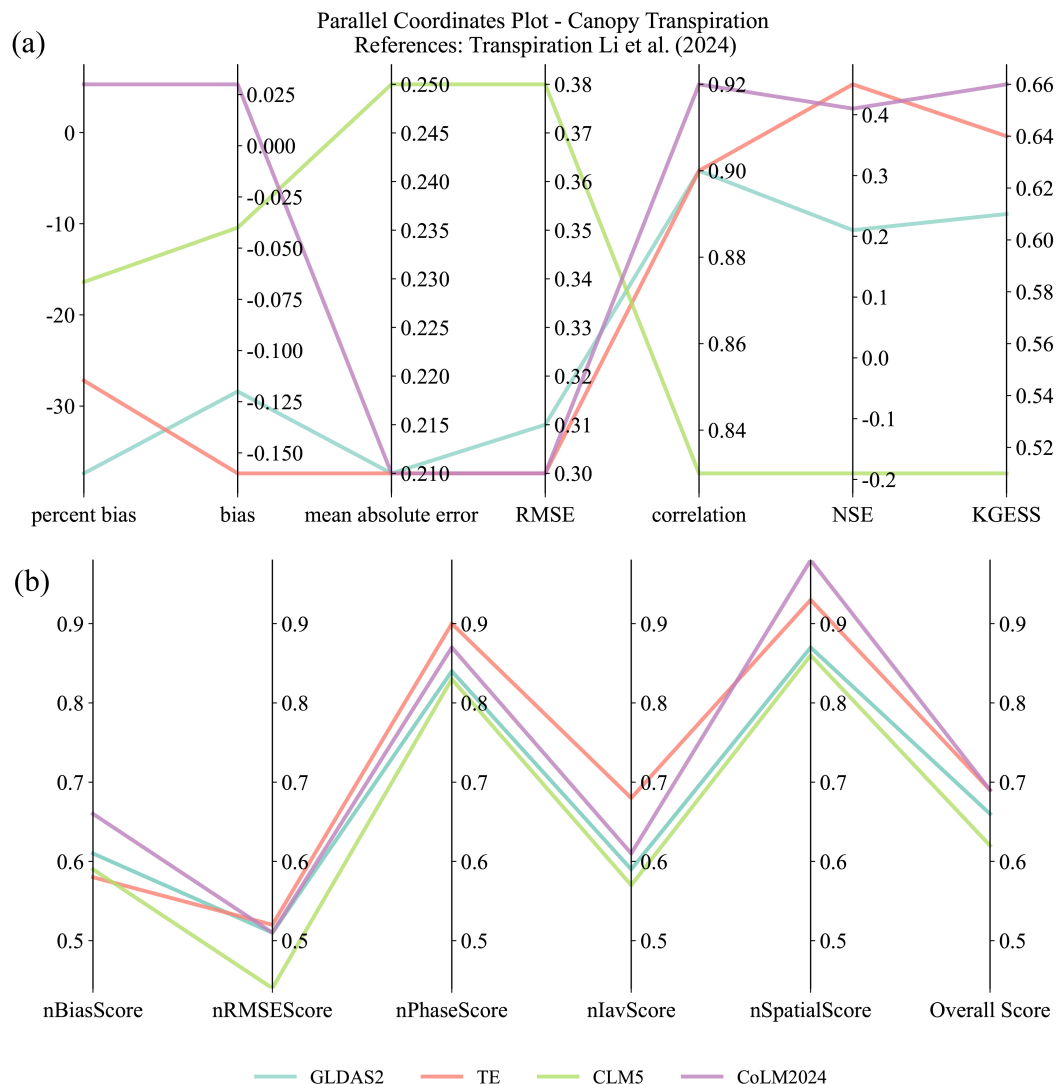



Figure 7: Evaluation of canopy transpiration using various (a) metrics and (b) score indexes for GLDAS2, TE, CLM5, and CoLM2024.

To demonstrate the multiple models' analytical capabilities of OpenBench, we conducted a comparative analysis of four state-of-the-art land surface models: GLDAS2, TE, CLM5, and CoLM2024. The evaluation period spanned from 2002 to 2006, utilizing a monthly temporal resolution. Multiple reference datasets were incorporated, including Li et al. (2024) for canopy transpiration, the FLUXCOM dataset (from ILAMB) and GLEAM4.2a for surface heat fluxes, and LORA (from ILAMB) for total runoff assessment (Hobeichi et al., 2019).

Figure 6 illustrates the comparative analysis through two complementary visualization approaches: a heat map and a parallel coordinates plot. The heat map (left panel) provides an intuitive visualization of relative model performance across different

variables, while the parallel coordinates plot (right panel) reveals intricate relationships between various performance metrics. This dual visualization strategy enables researchers to quickly identify patterns and trade-offs in model performance across multiple variables simultaneously. The analysis reveals that under current configurations, CoLM2024 and TE achieve the highest score for canopy transpiration, while CLM5 and CoLM2024 show the highest score for total runoff. CoLM2024 maintains relatively higher score in other variables.

For detailed variable-specific analysis, **Figure 7** presents an in-depth examination of canopy transpiration across all models, utilizing both conventional metrics (**Fig. 7a**) and normalized scores (**Fig. 7b**). The metrics analysis reveals that CoLM2024 exhibits a tendency to overestimate canopy transpiration, while other models show varying degrees of underestimation, as indicated by the percent bias metric. Furthermore, CoLM2024 achieves optimal performance regarding RMSE minimization, correlation maximization, and KGESS optimization. TE demonstrates particularly strong performance in NSE and ranks second in KGESS. Regarding scoring indices, TE excels in nRMSEScore, nPhaseScore, and nIavScore, whereas CoLM2024 achieves the highest nBiasScore.

This comprehensive comparative analysis not only highlights the relative strengths and weaknesses of each model but also offers valuable insights into their simulation capabilities regarding various aspects of land surface processes. Such a detailed evaluation helps identify areas where models excel or need further refinement, effectively guiding future development efforts.

4.2.2 Multiple parameterizations and multiple references

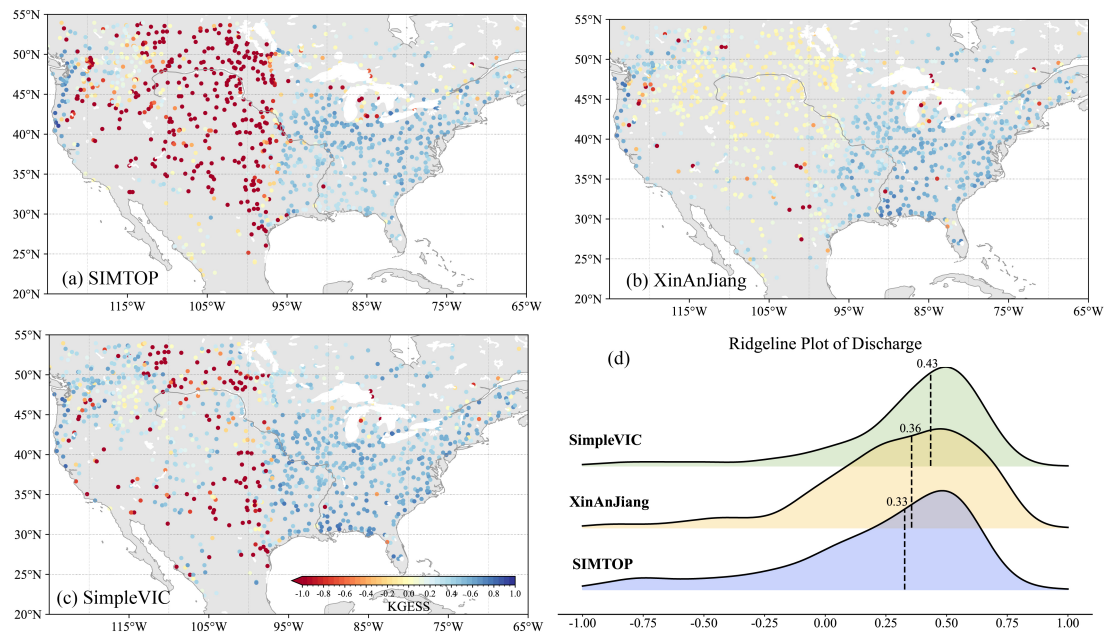



Figure 8: Comparison of daily discharge simulated from different parameterizations of the CoLM2024 model's runoff generation scheme with GRDC observations using KGESS metric.

To evaluate the versatility of OpenBench in analyzing model parameterization schemes, we conducted a comprehensive assessment of different runoff generation parameterizations within the CoLM2024 model framework. The analysis focused on daily discharge simulations at 0.1° resolution from 1985 to 1999, comparing three distinct parameterization approaches: SIMTOP, XinAnjiang, and SimpleVIC schemes. These simulations were evaluated against observational data from GRDC.

Figure 8 presents a spatial analysis of model performance using the KGESS metric across the continental United States. The station-based visualization (**Figs. 8a-c**) reveals distinct spatial patterns in model performance for each parameterization scheme. The SimpleVIC parameterization demonstrates superior performance across most regions, particularly in areas with complex hydrological processes. In contrast, the XinAnJiang scheme exhibits notable strengths in simulating discharge patterns within arid and semi-arid regions, suggesting its particular effectiveness in water-limited environments.

To further elucidate the statistical characteristics of these parameterizations, we employed a ridgeline plot analysis (**Fig. 8d**). This visualization technique effectively captures the distribution of performance metrics across different schemes, with the dashed lines and accompanying numbers indicating median values for each parameterization. The analysis confirms that the SimpleVIC parameterization achieves the highest overall performance metrics, though each scheme shows specific regional strengths.

OpenBench's capability to handle multiple reference datasets is demonstrated through a detailed evaluation of latent heat simulations. **Figure 9** illustrates this multi-reference analysis framework, comparing CoLM2024 simulations against four distinct reference sources: satellite-derived products (CLASS), machine learning outputs (FLUXCOM), in-situ measurements (PLUMBER2) (Ukkola et al., 2022), and reanalysis data (ERA5Land). This comparison was conducted at a monthly temporal resolution and 0.5° spatial resolution for the period 2002-2006. The resulting heat map visualization reveals strong model-data agreement across all reference datasets, with correlation coefficients consistently exceeding 0.90.

This comprehensive evaluation approach validates the model's performance against multiple independent data sources, as well as provides insights into the structural uncertainties inherent in different observational datasets. Such multi-reference validation is particularly valuable for variables where direct measurements are sparse and each observational approach has its own uncertainties and biases. The consistently high correlation values across different reference datasets enhance confidence in the model's ability to capture fundamental physical processes while also highlighting areas where uncertainties in observational data may impact validation efforts.

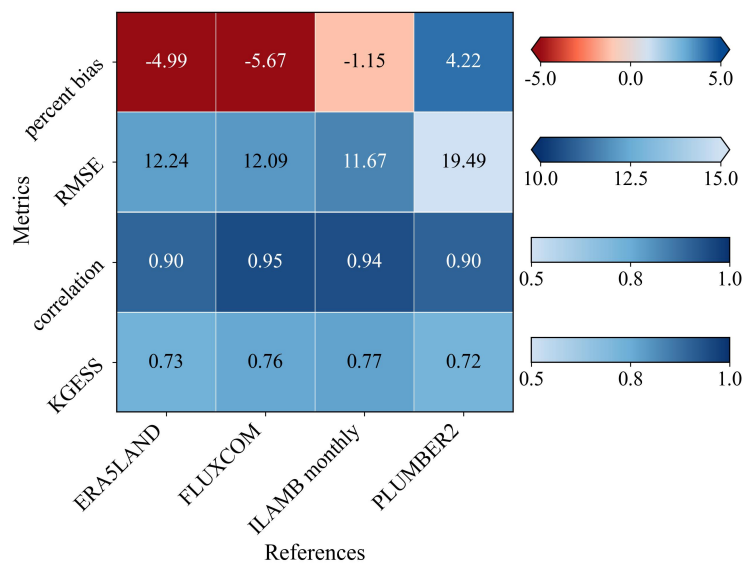
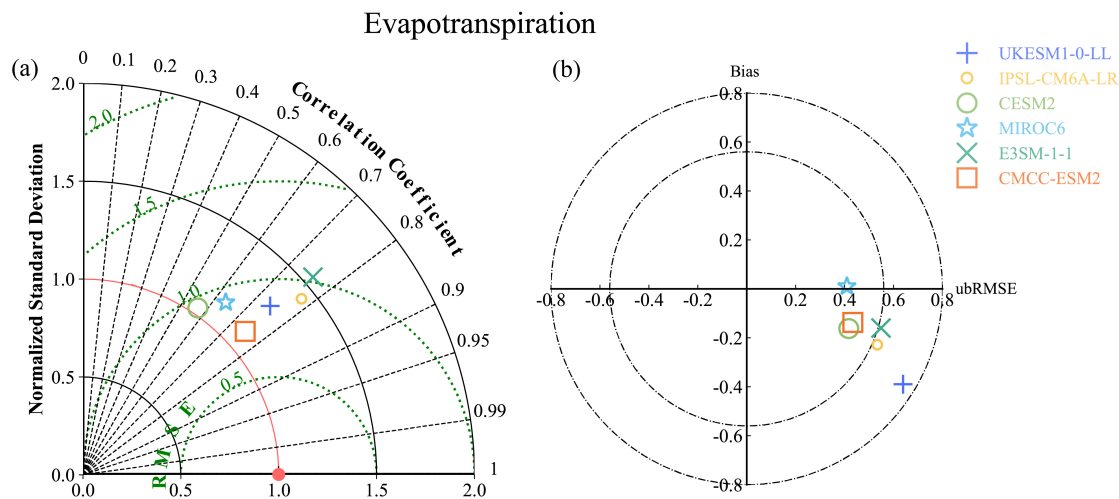



Figure 9: Evaluation of latent heat flux simulated by CoLM2024 using various metrics with different reference datasets.

4.2.3 CMIP styles comparison

Evapotranspiration

 Figure 10: The Taylor (a) and target diagram (b) for comparing evapotranspiration among the six models in LS3MIP.

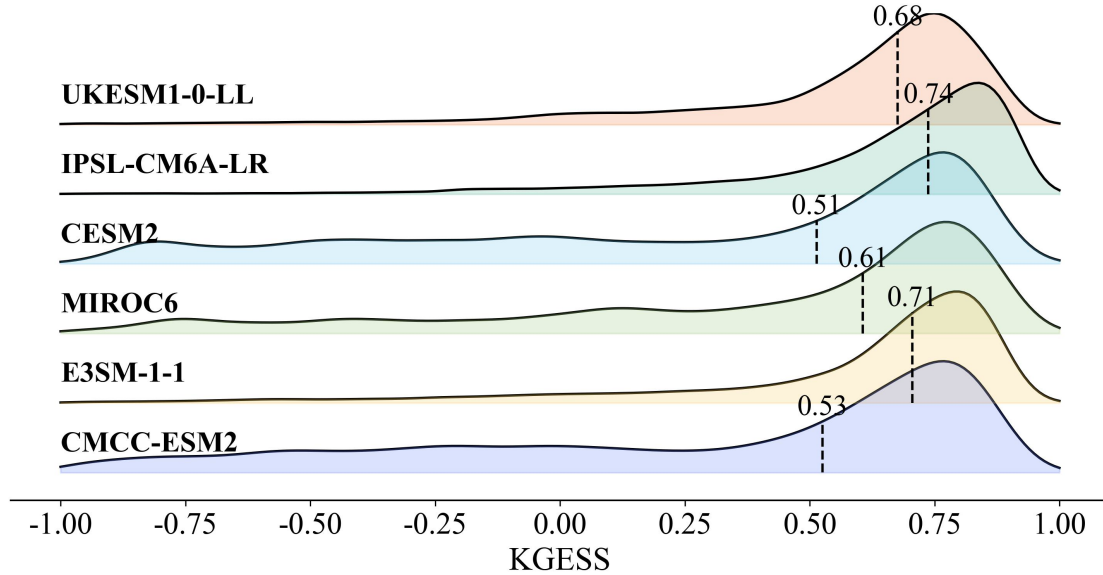



Figure 11: The ridgeline plot comparing evapotranspiration for the six models in LS3MIP Land-hist experiment.

OpenBench's evaluation framework incorporates robust capabilities for analyzing CMIP-style datasets, with particular emphasis on experimental outputs from initiatives such as ISIMIP and LS3MIP. The system's architecture includes specialized data processing modules designed to handle the standardized conventions of CMIP outputs, including variable naming conventions, temporal frequencies, and grid structures, ensuring seamless integration with the evaluation framework.

Figure 10 demonstrates this capability through a comprehensive analysis of evapotranspiration simulations from the LS3MIP experiment. The analysis employs both Taylor and target diagrams to provide complementary perspectives on model performance. The Taylor diagram (**Fig. 10a**) effectively visualizes the relationship between correlation coefficients, normalized standard deviations, and Centralized Root Mean Square Error. This multi-metric representation enables immediate identification of models that achieve optimal balance across these key performance indicators. The target diagram (**Fig. 10b**) supplements this analysis by providing additional insight into bias components and pattern variations, with distinct symbols differentiating between the various LS3MIP simulations.

To further elucidate the performance distribution across different models, **Figure 11** presents a ridgeline plot analysis of the KGESS metric. This visualization technique reveals the full spectrum of model performance, highlighting both central tendencies and variations in simulation quality. The analysis demonstrates that while certain models consistently achieve higher performance metrics, considerable variation in simulation quality exists across the ensemble. This variation provides valuable insights into the structural uncertainties inherent in current land surface modeling approaches.

The integration of CMIP-style evaluation capabilities within OpenBench serves multiple critical functions in the broader context of Earth system modeling. First, it enables systematic assessment of land surface processes within coupled climate

models, providing essential feedback for model development and improvement. Second, it facilitates direct comparisons between offline land surface model simulations and their behavior within coupled frameworks, helping to identify potential interactions and feedback that may affect model performance. Finally, this capability supports comprehensive model
550 intercomparison studies, contributing to our understanding of model uncertainties and their implications for future climate projections.

This robust framework for evaluating CMIP-style outputs positions OpenBench as a valuable tool for both model development and climate change research. By providing standardized, comprehensive evaluation metrics for these complex datasets, OpenBench supports the ongoing effort to improve our understanding and prediction of land surface processes in the context
555 of global climate change.

5 Extensibility and Customization

OpenBench is engineered with extensibility and customization as core design principles, enabling the system to evolve alongside the rapidly advancing field of land surface science. This flexible architecture accommodates the integration of new models, variables, datasets, measurement units, evaluation metrics, and scoring systems while maintaining operational
560 consistency and scientific rigor. The system's modular design facilitates seamless incorporation of new reference datasets through a streamlined configuration process. Researchers can integrate additional observational or reanalysis data by creating appropriate entries in the reference configuration file, specifying dataset locations and characteristics. This process involves defining dataset properties, including directory structures, temporal and spatial resolutions, and variable-specific parameters. For datasets with unique characteristics, users can develop custom processing scripts that integrate smoothly with the existing
565 evaluation framework.

Variable integration follows a similarly structured approach. Adding new variables requires coordinated updates to both reference and simulation configuration files, alongside corresponding dataset configurations that define variable properties. This process may include the development of specialized evaluation metrics and visualization components to effectively represent and analyze the new variables within the system's analytical framework. The integration of new land surface models
570 demonstrates OpenBench's architectural flexibility. Users can incorporate additional models by creating model-specific namelist files that establish straightforward mappings between model outputs and OpenBench's standardized variables. This integration is supported by updates to the simulation configuration and, where necessary, the development of custom variable filtering scripts to handle model-specific output characteristics. OpenBench's unit conversion system exemplifies its sophisticated approach to extensibility. The unit processing module employs a flexible design that readily accommodates new
575 measurement units for existing and new variables. Users can implement additional unit conversions by creating methods within the designated class, following established naming conventions. The system's dynamic method calling architecture ensures that new unit conversions integrate seamlessly into the evaluation workflow without requiring modifications to other system components. The system's evaluation framework maintains equal flexibility in incorporating new metrics and scoring

methodologies. Users can implement additional evaluation metrics by creating new methods within the metrics class, properly handling missing data, and maintaining comprehensive documentation. Similarly, new scoring systems can be integrated into the scores class within Mod_Scores.py, with appropriate attention to normalization procedures and interpretation guidelines. This comprehensive approach to extensibility guarantees that OpenBench stays at the forefront of land surface model evaluation capabilities. As new scientific questions emerge, new models are developed, and new observational datasets become available, the system can readily adapt to incorporate these advances. This flexibility is essential for maintaining a state-of-the-art evaluation framework that effectively serves the evolving needs of the land surface modeling community while ensuring consistent, high-quality analysis across various applications and research contexts.

6 Conclusions

Our newly developed OpenBench represents a significant advancement in land surface model evaluation methodology, addressing critical gaps in existing evaluation frameworks while introducing innovative capabilities for comprehensive model assessment. By integrating high-resolution benchmark datasets, sophisticated evaluation metrics, and efficient data handling mechanisms, OpenBench provides users with a powerful tool for enhancing the understanding and performance of land surface models. The system's key strengths lie in several areas. First, its ability to handle diverse data types and formats, from station-based measurements to gridded products, enables comprehensive evaluation across multiple spatial and temporal scales. Second, incorporating human activity impacts into the evaluation framework fills a crucial gap in current assessment tools, allowing for a more realistic evaluation of model performance in anthropogenically modified landscapes. Third, the system's robust computational architecture, built on efficient parallel processing and standardized data handling protocols, ensures scalability and reliability in processing large-scale datasets. The case studies presented demonstrate OpenBench's practical utility across various applications. The system has proven effective in identifying model strengths and areas requiring improvement, from evaluating hydrological processes and urban heat fluxes to assessing agricultural modeling capabilities. The multi-reference approach to model evaluation provides particularly valuable insights, helping distinguish between model deficiencies and observational uncertainties. OpenBench's extensible architecture ensures its continued relevance as the field evolves. The system's ability to incorporate new models, variables, datasets, and evaluation metrics allows it to adapt to emerging research needs and technological advances. This flexibility, combined with its comprehensive evaluation capabilities, positions OpenBench as a valuable resource for both model development and operational applications. Looking forward, OpenBench's role in advancing land surface modeling extends beyond technical evaluation. By providing standardized and reproducible evaluation methods, OpenBench facilitates more effective collaboration within the modeling community and supports more informed decision-making in environmental management. As we face increasing environmental challenges and seek to improve our understanding of Earth system processes, tools like OpenBench will be crucial in developing more accurate and reliable land surface models.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (under Grants 42475172, 42088101, 42075158, 42175158, 42375166, 42077168 and 42375164), the Guangdong Major Project of Basic and Applied Basic Research (2021B0301030007) and Guangdong Basic and Applied Basic Research Foundation (2024A1515010283). It is also supported by the National Key Scientific and Technological Infrastructure project "Earth System Science Numerical Simulator Facility" (EarthLab), and the specific research fund of The Innovation Platform for Academicians of Hainan Province (YSPTZX202143). Zhongwang Wei is supported by Guangdong Pearl River Talent Program (Young Talents) No. 2021QN02G307. We also acknowledge the high-performance computing support from the School of Atmospheric Science at Sun Yat-sen University.

Open Research

All codes and data used can be found in Wei (2025). The CoLM2024 model used in this study can be downloaded from https://github.com/CoLM-SYSU/CoLM202X. The high-resolution land surface characteristics data sets for CoLM2024 can be downloaded from http://globalchange.bnu.edu.cn/research/data. The OpenBench software is available at https://zenodo.org/records/15811122 and is updated routinely at https://github.com/zhongwangwei/OpenBench. The TE dataset is available at https://www.eorc.jaxa.jp/water/index.html; CLM5 dataset is available at https://rda.ucar.edu/datasets/d651011/; GLDAS2 dataset is available at https://ldas.gsfc.nasa.gov/gldas; LS3MIP dataset is available for downloading on the ESGF node: https://esgf-node.llnl.gov/search/cmip6/; The GRDC discharge datasets are available at https://grdc.bafg.de/GRDC.

Author contributions

ZW prepared the data, developed the models, analysed the results, and prepared the draft manuscript with the contributions of all co-authors. QX analysed the results, visualised the result, developed the model, tested the model and prepared the draft manuscript. FB developed the model, tested the model and edited the manuscript. XX and ZW tested the model and edited the manuscript. WD, HL, NW, XL, SZ and HY were both contributed to develop and test the models. LL and YD edited the manuscript.

Competing interests

The contact author has declared that none of the authors has any competing interests.

References

Abbaspour, K. C., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., Zobrist, J., and Srinivasan, R.: Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT, Journal of Hydrology, 333, 413-430, https://doi.org/10.1016/j.jhydrol.2006.09.014, 2007.

640 Appel, K. W., Gilliam, R. C., Davis, N., Zubrow, A., and Howard, S. C.: Overview of the atmospheric model evaluation tool (AMET) v1.1 for evaluating meteorological and air quality models, Environmental Modelling & Software, 26, 434-443, https://doi.org/10.1016/j.envsoft.2010.09.007, 2011.

Archdeacon, T. J.: Correlation and regression analysis: a historian's guide, Univ of Wisconsin Press1994.

Arora, V. K., Seiler, C., Wang, L., and Kou-Giesbrecht, S.: Towards an ensemble-based evaluation of land surface models in

645 light of uncertain forcings and observations, Biogeosciences, 20, 1313-1355, 2023.

Bai, F., Wei, Z., Wei, N., Lu, X., Yuan, H., Zhang, S., Liu, S., Zhang, Y., Li, X., and Dai, Y.: Global Assessment of Atmospheric Forcing Uncertainties in The Common Land Model 2024 Simulations, Journal of Geophysical Research: Atmospheres, 129, e2024JD041520, https://doi.org/10.1029/2024JD041520, 2024.

Barrett, J. P.: The coefficient of determination—some limitations, The American Statistician, 28, 19-20, 1974.

650 Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., van den Hurk, B. J. J., Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, Journal of Hydrometeorology, 16, 1425-1442, https://doi.org/10.1175/JHM-D-14-0158.1, 2015.

Blyth, E. M., Arora, V. K., Clark, D. B., Dadson, S. J., De Kauwe, M. G., Lawrence, D. M., Melton, J. R., Pongratz, J., Turton,

655 R. H., Yoshimura, K., and Yuan, H.: Advances in Land Surface Modelling, Current Climate Change Reports, 7, 45-71, 10.1007/s40641-021-00171-5, 2021.

Box, G. E.: Use and abuse of regression, Technometrics, 8, 625-629, 1966.

Chen, S., Li, L., Wei, Z., Wei, N., Zhang, Y., Zhang, S., Yuan, H., Shangguan, W., Zhang, S., Li, Q., and Dai, Y.: Exploring Topography Downscaling Methods for Hyper-Resolution Land Surface Modeling, Journal of Geophysical Research:

660 Atmospheres, 129, e2024JD041338, https://doi.org/10.1029/2024JD041338, 2024.

Cole, N. S.: Bias in testing, American Psychologist, 36, 1067, 1981.

Collier, N., Hoffman, F. M., Lawrence, D. M., Keppel-Aleks, G., Koven, C. D., Riley, W. J., Mu, M., and Randerson, J. T.: The International Land Model Benchmarking (ILAMB) System: Design, Theory, and Implementation, Journal of Advances in Modeling Earth Systems, 10, 2731-2754, https://doi.org/10.1029/2018MS001354, 2018.

665 Criss, R. E. and Winston, W. E.: Do Nash values have value? Discussion and alternate proposals, Hydrological Processes: An International Journal, 22, 2723-2725, 2008.

Dai, Y., Zeng, X., Dickinson, R. E., Baker, I., Bonan, G. B., Bosilovich, M. G., Denning, A. S., Dirmeyer, P. A., Houser, P. R., Niu, G., Oleson, K. W., Schlosser, C. A., and Yang, Z.-L.: The Common Land Model, Bulletin of the American Meteorological Society, 84, 1013-1024, https://doi.org/10.1175/BAMS-84-8-1013, 2003.

670 Dawson, C., Mount, N. J., Abrahart, R. J., and Shamseldin, A. Y.: Ideal point error for model assessment in data-driven river flow forecasting, Hydrology and Earth System Sciences, 16, 3049-3060, 2012.

Entekhabi, D., Reichle, R. H., Koster, R. D., and Crow, W. T.: Performance Metrics for Soil Moisture Retrievals and Application Requirements, Journal of Hydrometeorology, 11, 832-840, https://doi.org/10.1175/2010JHM1223.1, 2010.

Everitt, B. S.: The Cambridge dictionary of statistics, Cambridge2009.

675 Fan, H., Xu, Q., Bai, F., Wei, Z., Zhang, Y., Lu, X., Wei, N., Zhang, S., Yuan, H., Liu, S., Li, X., Li, X., and Dai, Y.: An Unstructured Mesh Generation Tool for Efficient High-Resolution Representation of Spatial Heterogeneity in Land Surface Models, Geophysical Research Letters, 51, e2023GL107059, https://doi.org/10.1029/2023GL107059, 2024.

Fisher, R. A. and Koven, C. D.: Perspectives on the Future of Land Surface Models and the Challenges of Representing Complex Terrestrial Systems, Journal of Advances in Modeling Earth Systems, 12, e2018MS001453,

680 https://doi.org/10.1029/2018MS001453, 2020.

Garcia, F., Folton, N., and Oudin, L.: Which objective function to calibrate rainfall–runoff models for low-flow index simulations?, Hydrological Sciences Journal, 62, 1149-1166, 10.1080/02626667.2017.1308511, 2017.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80-91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

685   Hamman, J. J., Nijssen, B., Bohn, T. J., Gergel, D. R., and Mao, Y.: The Variable Infiltration Capacity model version 5 (VIC-5): Infrastructure improvements for new applications and reproducibility, Geoscientific Model Development, 11, 3481-3496, 2018.

Hanazaki, R., Yamazaki, D., and Yoshimura, K.: Development of a Reservoir Flood Control Scheme for Global Flood Models, 690   Journal of Advances in Modeling Earth Systems, 14, e2021MS002944, https://doi.org/10.1029/2021MS002944, 2022.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, Nature, 585, 357-362, 10.1038/s41586-020-2649-2, 2020.

695   Haverd, V., Smith, B., Nieradzik, L., Briggs, P. R., Woodgate, W., Trudinger, C. M., Canadell, J. G., and Cuntz, M.: A new version of the CABLE land surface model (Subversion revision r4601) incorporating land use and land cover change, woody vegetation demography, and a novel optimisation-based approach to plant coordination of photosynthesis, Geosci. Model Dev., 11, 2995-3026, 10.5194/gmd-11-2995-2018, 2018.

He, C., Valayamkunnath, P., Barlage, M., Chen, F., Gochis, D., Cabell, R., Schneider, T., Rasmussen, R., Niu, G.-Y., and 700   Yang, Z.-L.: Modernizing the open-source community Noah with multi-parameterization options (Noah-MP) land surface model (version 5.0) with enhanced modularity, interoperability, and applicability, Geoscientific Model Development, 16, 5131-5151, 2023.

Hobeichi, S., Abramowitz, G., Evans, J., and Beck, H. E.: Linear Optimal Runoff Aggregate (LORA): A global gridded synthesis runoff product, Hydrology and Earth System Sciences, 23, 851-870, 2019.

705   Hoyer, S. and Hamman, J.: xarray: ND labeled arrays and datasets in Python, Journal of Open Research Software, 5, 10-10, 2017.

Hunter, J. D.: Matplotlib: A 2D Graphics Environment, 10.1109/mcse.2007.55, 2007.

Iizumi, T. and Sakai, T.: The global dataset of historical yields for major crops 1981–2016, Scientific Data, 7, 97, 10.1038/s41597-020-0433-7, 2020.

710   Jennings, K. S., Winchell, T. S., Livneh, B., and Molotch, N. P.: Spatial variation of the rain–snow temperature threshold across the Northern Hemisphere, Nature Communications, 9, 1148, 10.1038/s41467-018-03629-7, 2018.

Jensen, T., J. Prestopnik, H. Soh, L. Goodrich, B. Brown, R. Bullock, J. Halley Gotway, K. Newman, J. Opatz,: The MET Version 11.1.1 User's Guide, Developmental Testbed Center, 2024.

Joblib, D. T.: Joblib: running Python functions as pipeline jobs, 2020.

715   Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., Papale, D., Schwalm, C., Tramontana, G., and Reichstein, M.: The FLUXCOM ensemble of global land-atmosphere energy fluxes, Scientific Data, 6, 74, 10.1038/s41597-019-0076-8, 2019.

Kenney, J. F. and Keeping, E.: Root mean square, Mathematics of statistics, 1, 59-60, 1962.

Knoben, W. J., Freer, J. E., and Woods, R. A.: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta 720   efficiency scores, Hydrology and Earth System Sciences, 23, 4323-4331, 2019.

Krause, P., Boyle, D., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Advances in geosciences, 5, 89-97, 2005.

Krstic, G., Krstic, N. S., and Zambrano-Bigiarini, M.: The br2–weighting Method for Estimating the Effects of Air Pollution on Population Health, Journal of Modern Applied Statistical Methods, 15, 42, 2016.

725   Kumar, S. V., Peters-Lidard, C. D., Santanello, J., Harrison, K., Liu, Y., and Shaw, M.: Land surface Verification Toolkit (LVT)–a generalized framework for land surface model evaluation, Geoscientific Model Development, 5, 869-886, 2012.

Lamontagne, J. R., Barber, C. A., and Vogel, R. M.: Improved Estimators of Model Performance Efficiency for Skewed Hydrologic Data, Water Resources Research, 56, e2020WR027101, https://doi.org/10.1029/2020WR027101, 2020.

Lawrence, D. M., Fisher, R. A., Koven, C. D., Oleson, K. W., Swenson, S. C., Bonan, G., Collier, N., Ghimire, B., van 730   Kampenhout, L., Kennedy, D., Kluzek, E., Lawrence, P. J., Li, F., Li, H., Lombardozzi, D., Riley, W. J., Sacks, W. J., Shi, M., Vertenstein, M., Wieder, W. R., Xu, C., Ali, A. A., Badger, A. M., Bisht, G., van den Broeke, M., Brunke, M. A., Burns, S. P., Buzan, J., Clark, M., Craig, A., Dahlin, K., Drewniak, B., Fisher, J. B., Flanner, M., Fox, A. M., Gentine, P., Hoffman, F.,

Keppel-Aleks, G., Knox, R., Kumar, S., Lenaerts, J., Leung, L. R., Lipscomb, W. H., Lu, Y., Pandey, A., Pelletier, J. D., Perket, J., Randerson, J. T., Ricciuto, D. M., Sanderson, B. M., Slater, A., Subin, Z. M., Tang, J., Thomas, R. Q., Val Martin, M., and Zeng, X.: The Community Land Model Version 5: Description of New Features, Benchmarking, and Impact of Forcing Uncertainty, Journal of Advances in Modeling Earth Systems, 11, 4245-4287, https://doi.org/10.1029/2018MS001583, 2019.

Lee, J., Gleckler, P. J., Ahn, M.-S., Ordonez, A., Ullrich, P. A., Sperber, K. R., Taylor, K. E., Planton, Y. Y., Guilyardi, E., and Durack, P.: Objective evaluation of earth system models: PCMDI metrics package (PMP) version 3, EGUsphere, 2023, 1-48, 2023.

Legates, D. R. and McCabe Jr., G. J.: Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation, Water Resources Research, 35, 233-241, https://doi.org/10.1029/1998WR900018, 1999.

Li, C., Han, J., Liu, Z., Tu, Z., and Yang, H.: A harmonized global gridded transpiration product based on collocation analysis, Scientific Data, 11, 604, 10.1038/s41597-024-03425-7, 2024.

Li, D., Ji, J., and Li, Y.: Climate simulations based on a different-grid nested and coupled model, Advances in Atmospheric Sciences, 19, 487-499, 10.1007/s00376-002-0081-1, 2002.

Liang, X.-Z., Xu, M., Yuan, X., Ling, T., Choi, H. I., Zhang, F., Chen, L., Liu, S., Su, S., Qiao, F., He, Y., Wang, J. X. L., Kunkel, K. E., Gao, W., Joseph, E., Morris, V., Yu, T.-W., Dudhia, J., and Michalakes, J.: Regional Climate–Weather Research and Forecasting Model, Bulletin of the American Meteorological Society, 93, 1363-1387, https://doi.org/10.1175/BAMS-D-11-00180.1, 2012.

Liu, L., Fisher, R. A., Douville, H., Padrón, R. S., Berg, A., Mao, J., Alessandri, A., Kim, H., and Seneviratne, S. I.: No constraint on long-term tropical land carbon-climate feedback uncertainties from interannual variability, Communications Earth & Environment, 5, 348, 10.1038/s43247-024-01504-6, 2024.

Lo, J. C.-F., Yang, Z.-L., and Pielke Sr., R. A.: Assessment of three dynamical climate downscaling methods using the Weather Research and Forecasting (WRF) model, Journal of Geophysical Research: Atmospheres, 113, https://doi.org/10.1029/2007JD009216, 2008.

Lovie, P.: Coefficient of Variation, in: Encyclopedia of Statistics in Behavioral Science, https://doi.org/10.1002/0470013192.bsa107, 2005.

Makridakis, S., Andersen, A., Carbone, R., Fildes, R., Hibon, M., Lewandowski, R., Newton, J., Parzen, E., and Winkler, R.: The accuracy of extrapolation (time series) methods: Results of a forecasting competition, Journal of Forecasting, 1, 111-153, https://doi.org/10.1002/for.3980010202, 1982.

MANABE, S.: CLIMATE AND THE OCEAN CIRCULATION: I. THE ATMOSPHERIC CIRCULATION AND THE HYDROLOGY OF THE EARTH'S SURFACE, Monthly Weather Review, 97, 739-774, https://doi.org/10.1175/1520-0493(1969)097<0739:CATOC>2.3.CO;2, 1969.

McKinney, W.: Data structures for statistical computing in Python, SciPy, 445, 51-56, 2010.

Met Office, U.: Cartopy: A cartographic python library with a matplotlib interface, Exeter, Devon, 2010.

Miralles, D. G., Holmes, T., De Jeu, R., Gash, J., Meesters, A., and Dolman, A.: Global land-surface evaporation estimated from satellite-based observations, Hydrology and Earth System Sciences, 15, 453-469, 2011.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., and Kumar, R.: On the choice of calibration metrics for "high-flow" estimation using hydrologic models, Hydrology and Earth System Sciences, 23, 2601-2614, 2019.

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., and Hersbach, H.: ERA5-Land: A state-of-the-art global reanalysis dataset for land applications, Earth system science data, 13, 4349-4383, 2021.

Myung, I. J.: Tutorial on maximum likelihood estimation, Journal of Mathematical Psychology, 47, 90-100, https://doi.org/10.1016/S0022-2496(02)00028-7, 2003.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I — A discussion of principles, Journal of Hydrology, 10, 282-290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nelson, J. A., Walther, S., Gans, F., Kraft, B., Weber, U., Novick, K., Buchmann, N., Migliavacca, M., Wohlfahrt, G., and Šigut, L.: X-BASE: the first terrestrial carbon and water flux products from an extended data-driven scaling framework, FLUXCOM-X, Biogeosciences, 21, 5079-5115, 2024.

Nitta, T., Yoshimura, K., Takata, K., O'ishi, R., Sueyoshi, T., Kanae, S., Oki, T., Abe-Ouchi, A., and Liston, G. E.: Representing Variability in Subgrid Snow Cover and Snow Depth in a Global Land Model: Offline Validation, Journal of Climate, 27, 3318-3330, https://doi.org/10.1175/JCLI-D-13-00310.1, 2014.

785 Olkin, I. and Pratt, J. W.: Unbiased Estimation of Certain Correlation Coefficients, The Annals of Mathematical Statistics, 29, 201-211, 1958.

Pearson, K.: Notes on the History of Correlation, Biometrika, 13, 25-45, 10.2307/2331722, 1920.

Phillips, A. S., Deser, C., and Fasullo, J.: Evaluating Modes of Variability in Climate Models, Eos, Transactions American Geophysical Union, 95, 453-455, https://doi.org/10.1002/2014EO490002, 2014.

Pizarro, A. and Jorquera, J.: Advancing objective functions in hydrological modelling: Integrating knowable moments for 790 improved simulation accuracy, Journal of Hydrology, 634, 131071, https://doi.org/10.1016/j.jhydrol.2024.131071, 2024.

Pokhrel, Y. N., Hanasaki, N., Wada, Y., and Kim, H.: Recent progresses in incorporating human land–water management into global land surface models toward their integration into Earth system models, WIREs Water, 3, 548-574, https://doi.org/10.1002/wat2.1150, 2016.

Prigent, C., Jimenez, C., and Bousquet, P.: Satellite-Derived Global Surface Water Extent and Dynamics Over the Last 795 25 Years (GIEMS-2), Journal of Geophysical Research: Atmospheres, 125, e2019JD030711, https://doi.org/10.1029/2019JD030711, 2020.

Rocklin, M.: Dask: Parallel computation with blocked algorithms and task scheduling, SciPy, 126-132,

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., Arsenault, K., Cosgrove, B., Radakovich, J., Bosilovich, M., Entin, J. K., Walker, J. P., Lohmann, D., and Toll, D.: The Global Land Data Assimilation System, Bulletin 800 of the American Meteorological Society, 85, 381-394, https://doi.org/10.1175/BAMS-85-3-381, 2004.

Schilperoort, B., Liu, Y., and Doore, K.: xarray-regrid, 2024.

Schulzweida, U.: CDO User Guide, 10.5281/zenodo.10020800, 2023.

Sorooshian, S., Duan, Q., and Gupta, V. K.: Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model, Water Resources Research, 29, 1185-1194, 805 https://doi.org/10.1029/92WR02617, 1993.

Spearman, C.: " General Intelligence" Objectively Determined and Measured, 1961.

Stephen, K. D. and Kazemi, A.: Improved normalization of time-lapse seismic data using normalized root mean square repeatability data to improve automatic production and seismic history matching in the Nelson field, Geophysical Prospecting, 62, 1009-1027, https://doi.org/10.1111/1365-2478.12109, 2014.

810 Ukkola, A. M., Abramowitz, G., and De Kauwe, M. G.: A flux tower dataset tailored for land model evaluation, ESSD, 14, 449–461-449–461, 2022.

Van den Hurk, B., Kim, H., Krinner, G., Seneviratne, S. I., Derksen, C., Oki, T., Douville, H., Colin, J., Ducharne, A., and Cheruy, F.: LS3MIP (v1. 0) contribution to CMIP6: the Land Surface, Snow and Soil moisture Model Intercomparison Project– aims, setup and expected outcome, Geoscientific Model Development, 9, 2809-2832, 2016.

815 Varquez, A. C. G., Kiyomoto, S., Khanh, D. N., and Kanda, M.: Global 1-km present and future hourly anthropogenic heat flux, Scientific Data, 8, 64, 10.1038/s41597-021-00850-w, 2021.

Verseghy, D. L.: Class—A Canadian land surface scheme for GCMS. I. Soil model, International Journal of Climatology, 11, 111-133, https://doi.org/10.1002/joc.3370110202, 1991.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, 820 W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Vijaykumar, A., Bardelli, A. P., Rothberg, A., Hilboll, A., Kloeckner, A., Scopatz, A., Lee, A., Rokem, A., Woods, C. N., Fulton, C., Masson, C., Häggström, C., Fitzgerald, C., Nicholson, D. A., Hagen, D. R., Pasechnik, D. V., Olivetti, E., Martin, E., Wieser, E., Silva, F., 825 Lenders, F., Wilhelm, F., Young, G., Price, G. A., Ingold, G.-L., Allen, G. E., Lee, G. R., Audren, H., Probst, I., Dietrich, J. P., Silterra, J., Webber, J. T., Slavič, J., Nothman, J., Buchner, J., Kulick, J., Schönberger, J. L., de Miranda Cardoso, J. V., Reimer, J., Harrington, J., Rodríguez, J. L. C., Nunez-Iglesias, J., Kuczynski, J., Tritz, K., Thoma, M., Newville, M., Kümmerer, M., Bolingbroke, M., Tartre, M., Pak, M., Smith, N. J., Nowaczyk, N., Shebanov, N., Pavlyk, O., Brodtkorb, P. A., Lee, P., McGibbon, R. T., Feldbauer, R., Lewis, S., Tygier, S., Sievert, S., Vigna, S., Peterson, S., More, S., Pudlik, T., Oshima, T., 830 Pingel, T. J., Robitaille, T. P., Spura, T., Jones, T. R., Cera, T., Leslie, T., Zito, T., Krauss, T., Upadhyay, U., Halchenko, Y.

O., Vázquez-Baeza, Y., and SciPy, C.: SciPy 1.0: fundamental algorithms for scientific computing in Python, Nature Methods, 17, 261-272, 10.1038/s41592-019-0686-2, 2020.

Ward, N. D., Megonigal, J. P., Bond-Lamberty, B., Bailey, V. L., Butman, D., Canuel, E. A., Diefenderfer, H., Ganju, N. K., Goñi, M. A., Graham, E. B., Hopkinson, C. S., Khangaonkar, T., Langley, J. A., McDowell, N. G., Myers-Pigg, A. N., Neumann, R. B., Osburn, C. L., Price, R. M., Rowland, J., Sengupta, A., Simard, M., Thornton, P. E., Tzortziou, M., Vargas, R., Weisenhorn, P. B., and Windham-Myers, L.: Representing the function and sensitivity of coastal interfaces in Earth system models, Nature Communications, 11, 2458, 10.1038/s41467-020-16236-2, 2020.

Wartenburger, R., Seneviratne, S. I., Hirschi, M., Chang, J., Ciais, P., Deryng, D., Elliott, J., Folberth, C., Gosling, S. N., Gudmundsson, L., Henrot, A.-J., Hickler, T., Ito, A., Khabarov, N., Kim, H., Leng, G., Liu, J., Liu, X., Masaki, Y., Morfopoulos, C., Müller, C., Schmied, H. M., Nishina, K., Orth, R., Pokhrel, Y., Pugh, T. A. M., Satoh, Y., Schaphoff, S., Schmid, E., Sheffield, J., Stacke, T., Steinkamp, J., Tang, Q., Thiery, W., Wada, Y., Wang, X., Weedon, G. P., Yang, H., and Zhou, T.: Evapotranspiration simulations in ISIMIP2a—Evaluation of spatio-temporal characteristics with a comprehensive ensemble of independent datasets, Environmental Research Letters, 13, 075001, 10.1088/1748-9326/aac4bb, 2018.

Wei, Z.: Cases data and high quality figures of "OpenBench: a land models evaluation system [dataset], https://doi.org/10.5281/zenodo.15608222, 2025.

Wei, Z., He, X., Zhang, Y., Pan, M., Sheffield, J., Peng, L., Yamazaki, D., Moiz, A., Liu, Y., and Ikeuchi, K.: Identification of uncertainty sources in quasi-global discharge and inundation simulations using satellite-based precipitation products, Journal of Hydrology, 589, 125180, https://doi.org/10.1016/j.jhydrol.2020.125180, 2020.

Weigel, K., Bock, L., Gier, B. K., Lauer, A., Righi, M., Schlund, M., Adeniyi, K., Andela, B., Arnone, E., and Berg, P.: Earth System Model Evaluation Tool (ESMValTool) v2. 0–diagnostics for extreme events, regional and impact evaluation and analysis of Earth system models in CMIP, Geoscientific Model Development Discussions, 2020, 1-43, 2020.

Willmott, C. J., Robeson, S. M., and Matsuura, K.: A refined index of model performance, International Journal of climatology, 32, 2088-2094, 2012.

Xu, Q., Li, L., Wei, Z., Lu, X., Wei, N., Lee, X., and Dai, Y.: A multimodal machine learning fused global 0.1° daily evapotranspiration dataset from 1950-2022, Agricultural and Forest Meteorology, 372, 110645, https://doi.org/10.1016/j.agrformet.2025.110645, 2025.

Yamazaki, D., de Almeida, G. A. M., and Bates, P. D.: Improving computational efficiency in global river models by implementing the local inertial flow equation and a vector-based river network map, Water Resources Research, 49, 7221-7235, https://doi.org/10.1002/wrcr.20552, 2013.

Yang, Y., Pan, M., Lin, P., Beck, H. E., Zeng, Z., Yamazaki, D., David, C. H., Lu, H., Yang, K., Hong, Y., and Wood, E. F.: Global Reach-Level 3-Hourly River Flood Reanalysis (1980–2019), Bulletin of the American Meteorological Society, 102, E2086-E2105, https://doi.org/10.1175/BAMS-D-20-0057.1, 2021.

Yapo, P. O., Gupta, H. V., and Sorooshian, S.: Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data, Journal of Hydrology, 181, 23-48, https://doi.org/10.1016/0022-1694(95)02918-4, 1996.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, https://doi.org/10.1029/2007WR006716, 2008.

Zambrano-Bigiarini, M. and Bellin, A.: Comparing Goodness-of-fit Measures for Calibration of Models Focused on Extreme Events, EGU General Assembly Conference Abstracts, 11549,

Zhang, M.-Z., Xu, Z., Han, Y., and Guo, W.: An improved multivariable integrated evaluation method and tool (MVIETool) v1. 0 for multimodel intercomparison, Geoscientific Model Development, 14, 3079-3094, 2021.

Zhou, J., Xia, J., Wei, N., Liu, Y., Bian, C., Bai, Y., and Luo, Y.: A traceability analysis system for model evaluation on land carbon dynamics: design and applications, Ecological Processes, 10, 12, 10.1186/s13717-021-00281-w, 2021.

Zhou, X., Yamazaki, D., Revel, M., Zhao, G., and Modi, P.: Benchmark framework for global river model (Version 1.0), Authorea Preprints, 2024.

Zhuang, J., raphael, d., Huard, D., Bourgault, P., Banihirwe, A., Raynaud, S., Malevich, B., Schupfner, M., Filipe, Levang, S., Jüling, A., Almansi, M., RichardScottOz, RondeauG, Rasp, S., Smith, T. J., Stachelek, J., Plough, M., Pierre, Bell, R., and Li, X.: pangeo-data/xESMF: v0.7.1, 2023.